Limit theorems for branching processes with mutations

Cécile Delaporte

Université Pierre et Marie Curie
ED 386 - Sciences mathématiques de Paris centre
Laboratoire de Probabilités et Modèles Aléatoires

Thèse de doctorat de mathématiques

Théorèmes limites pour les processus de branchement avec mutations

Cécile Delaporte

sous la direction d'Amaury Lambert

présentée et soutenue publiquement le 2 octobre 2014

Devant un jury composé de :

M. ACHAZ Guillaume, Maître de conférence, examinateur
M. BANSAYE Vincent, Professeur chargé de cours, examinateur
M. CHAUMONT Loïc, Professeur, examinateur
M. DUQUESNE Thomas, Professeur, examinateur
Mme HAAS Bénédicte, Maître de conférence, rapporteur
M. LAMBERT Amaury, Professeur, directeur

Table des matières

Introduction

In this thesis, we develop theoretical results on the genealogical and mutational patterns produced by branching population models called *splitting trees* [Gei96, GK97, Lam10]. These models are a generalization of the birth-death model, in which the lifetimes of individuals do not necessarily follow exponential distributions, and individuals might have infinitely many offspring.

In population genetics, the null model applied by biologists is the Kingman coalescent, which arises as a scaling limit of a broad class of constant population size models, such as the Wright-Fisher or Moran models. Mutations are then traditionally modeled by introducing a mutation probability $\mu = \theta/N$ per generation (and per lineage), where N is the population size, and by assuming that each individual undergoes a mutation with probability μ. These mutations are assumed to be neutral, meaning that they do not affect the population dynamics. Time is then measured in units of N generations, and in the scaling limit of the Kingman coalescent mutations are independent of the genealogy, occurring at rate θ on the lineages. The existence of a similar invariance principle for branching populations then arises as a natural question, which we focus on in the first two chapters of the thesis.
In these chapters, we study the behavior, in the large population asymptotic, of genealogies of splitting trees with mutations at birth, conditioned on survival at a fixed time horizon. The results we obtain are based on the study of excursions of bivariate Lévy processes called Lévy processes *with marks*, coding for the random trees we study through their contour processes. We first state in Chapter I some limit theorems for Lévy processes with marks. These technical results are then applied in Chapter II to establish the convergence in distribution of a point measure called *marked coalescent point process*, describing the genealogy enriched with the mutational history of the extant population.

In Chapter III, we are interested in the patterns of genetic diversity for samples in a particular class of critical branching populations. Numerous results on the allelic partition for splitting trees have been established earlier [Lam08, CL12a, CL12b, Ric14], but the models we study here differ from those in the aforementioned papers mainly through two aspects. On the one hand, we consider here samples of the extant population. On the other hand, various assumptions on the foundation time of the population are considered, namely first, a fixed time of origin, and second, a random time of origin (with different possible prior distributions). In particular, the sampling is essential for the model to be relevant, as emphasized in Chapter IV, where some perspectives concerning the application of our model to real data are developed.

The purpose of the present chapter is to make the understanding of the next chapters easier to the reader. To this aim we first introduce the mathematical framework we work with, along with objects and notions that we extensively use in the sequel. Then we provide a summary of the main results for Chapters I, II and III.

0.1 Preliminaries

0.1.1 Topology and notation

We consider the Euclidean space $\mathbb{R}^d$ and endow it with its Borel σ-field $\mathcal{B}(\mathbb{R}^d)$. For all $x \in \mathbb{R}^d$, ${}^t x$ will denote its transpose. We denote by $\mathbb{D}(\mathbb{R}^d)$ the space of all càd-làg functions from $\mathbb{R}_+$ to $\mathbb{R}^d$. We endow the latter with the Skorokhod topology, which makes it a Polish space (see [JS87, VI.1.b]). In the sequel, for any function $f \in \mathbb{D}(\mathbb{R})$ and $x > 0$, we will use the notation $\Delta f(x) = f(x) - f(x-)$, where $f(x-) = \lim_{u \to x,\, u<x} f(u)$.

Now for any Polish space E, with its Borel σ-field $\mathcal{B}(E)$, we denote by $\mathcal{M}(E)$ the space of positive σ-finite measures on $(E, \mathcal{B}(E))$, and by $\mathcal{M}_f(E)$ the space of positive finite measures on $(E, \mathcal{B}(E))$. The space $\mathcal{M}_f(E)$ can be endowed with the weak topology, i.e. the coarsest topology for which the mappings $\mu \mapsto \int f \mathrm{d}\mu$ are continuous for any continuous bounded function f. In the sequel, we will use the notation $\mu(f) := \int f \mathrm{d}\mu$.

Hence we endow here $\mathcal{M}_f(\mathbb{R}^d)$ and $\mathcal{M}_f(\mathbb{D}(\mathbb{R}^d))$ with their respective weak topologies. The notation $\Rightarrow$ will be used for both weak convergence in $\mathbb{R}^d$ and in $\mathbb{D}(\mathbb{R}^d)$, and we will use the symbol $\stackrel{\mathcal{L}}{=}$ for the equality in law. Recall that for any sequence of $\mathbb{R}^d$-valued càd-làg processes (X_n), the weak convergence of (X_n) towards a process X of $\mathbb{D}(\mathbb{R}^d)$ is equivalent to the finite dimensional convergence of (X_n) towards X along any dense subset $D \subset \mathbb{R}_+$, together with the tightness of (X_n). For more details about convergence in distribution in $\mathbb{D}(\mathbb{R}^d)$, see [JS87, VI.3].

0.1.2 Random measures

Definitions

We recall here some notions about random measures. Consider a Polish space E, endowed with its Borel σ-field $\mathcal{B}(E)$. We call *random measure* on E a random variable ξ with values in $\mathcal{M}(E)$, endowed with the σ-field generated by the set of maps $\{p_B, B \in \mathcal{B}(E)\}$, where for any $\mu \in \mathcal{M}(E)$, $p_B(\mu) = \mu(B)$. For any $B \in \mathcal{B}(E)$, $\xi(B)$ is then a random variable in $[0, \infty]$. If ξ is integer-valued, it is called a *random point measure* or *point process*. In this case, ξ can be written $\sum_{i \in I} \delta_{x_i}$, where I is a countable set, and for any $i \in I$, x_i is called an atom of ξ and is a random element in E. If all atoms are distinct, the measure ξ is said to be *simple*.

Characterization of the law of a random measure

Two random measures ξ and ξ' are equal in law (denoted by $\xi \stackrel{\mathcal{L}}{=} \xi'$) if and only if for any $k \in \mathbb{N}$ and any $B_1, \ldots, B_k \in \mathcal{B}(E)$, $(\xi(B_1), \ldots, \xi(B_k)) \stackrel{\mathcal{L}}{=} (\xi'(B_1), \ldots, \xi'(B_k))$. Moreover, if ξ is a simple point process, the previous criterion reduces to the following : $\xi \stackrel{\mathcal{L}}{=} \xi'$ iff for any $B \in \mathcal{B}(E)$, $\xi(B) \stackrel{\mathcal{L}}{=} \xi'(B)$.

Poisson random measures

Let $m \in \mathcal{M}(E)$. A random measure ξ is called *Poisson random measure* with intensity m if

(a) For any $B \in \mathcal{B}(E)$, $\xi(B)$ follows a Poisson distribution with parameter $m(B)$.

(b) For any $k \in \mathbb{N}$ and $B_1, \ldots, B_k$ pairwise disjoint sets in $\mathcal{B}(E)$, the random variables $\xi(B_1), \ldots, \xi(B_k)$ are independent.

Note that a Poisson measure is thus characterized by its intensity measure.

We now state the *restriction property* for Poisson random measure.

Proposition 0.1. *Fix $k \in \mathbb{N}$, and let $E_1, \dots, E_k \in \mathcal{B}(E)$ be a partition of E. Consider ξ a Poisson measure with intensity m and denote by ξ_i the restriction of ξ to E_i, i.e. for any $B \in \mathcal{B}(E)$, $\xi_i(B) = \xi(B \cap E_i)$. Then the measures ξ_i, $i \in \{1, \dots, k\}$ are independent Poisson measures with respective intensities $m\mathbb{1}_{E_i}$.*

Convergence of random measures

Let $\xi, \xi_1, \xi_2, \dots$ be random measures on $\mathbb{R}^d$. The σ-field in $\mathcal{M}(\mathbb{R}^d)$ generated by $\{p_B, B \in \mathcal{B}(\mathbb{R}^d)\}$ coincides with the σ-field generated by the projections $p_f : \mu \mapsto \mu(f)$, for all continuous functions $f : \mathbb{R}^d \to \mathbb{R}_+$ with compact support. The space $\mathcal{M}(\mathbb{R}^d)$ endowed with this σ-field is Polish, and the convergence in distribution of a sequence (ξ_n) towards ξ (denoted by $\xi_n \Rightarrow \xi$) is defined as the convergence of the sequence $\mathbb{E}(f(\xi_n))$ towards $\mathbb{E}(f(\xi))$ for any continuous function $f : \mathbb{R}^d \to \mathbb{R}_+$ with compact support. In the sequel we use the following **characterization of the convergence in distribution** [Kal02, Th.16.16] in the particular case of a simple point measure in the limit :

Proposition 0.2. *Let (ξ_n) be a sequence of random measures on $\mathbb{R}^d$ and ξ a simple point measure on $\mathbb{R}^d$. Then $\xi_n \Rightarrow \xi$ iff $\xi_n(B) \Rightarrow \xi(B)$ for any relatively compact set $B \in \mathcal{B}(\mathbb{R}^d)$, such that $\xi(\partial B) = 0$, where ∂B denotes the boundary of B.*

Finally, we recall here the **law of rare events for null arrays of point measures** [Kal02, Th.16.18]. This result is a point measure counterpart of the law of rare events for random variables, which states that the Poisson distribution arises as limiting distribution for the number of successes in a large number of i.i.d. trials with small success probability.
Let $(\xi_j^n)_{n \in \mathbb{N}, 1 \le j \le p_n}$ be a sequence of random measures on $\mathbb{R}^d$, with (p_n) a non decreasing sequence of integers satisfying $\lim_{n\to\infty} p_n = \infty$. The measures (ξ_j^n) are said to form a *null array* if

- for any $n \in \mathbb{N}$, the random measures ξ_j^n, $1 \le j \le p_n$, are independent,
- for any relatively compact set $B \in \mathcal{B}(\mathbb{R}^d)$, $\sup_j \mathbb{E}(\xi_j^n(B) \wedge 1) \underset{n\to\infty}{\longrightarrow} 0$.

Theorem 0.3. *[Kal02, Th.16.18] Consider (ξ_j^n) a null array of point measures on $\mathbb{R}^d$, and ξ a Poisson point measure on $\mathbb{R}^d$ with intensity measure $m \in \mathcal{M}(\mathbb{R}^d)$. Then the sequence of point measures $(\sum_j \xi_j^n)_n$ converges in distribution towards ξ iff the following conditions hold :*

(i) *for any relatively compact set $B \in \mathcal{B}(\mathbb{R}^d)$ satisfying $m(\partial B) = 0$, $\sum_j \mathbb{P}(\xi_j^n(B) > 0) \underset{n\to\infty}{\longrightarrow} m(B)$,*

(ii) *for any relatively compact set $B \in \mathcal{B}(\mathbb{R}^d)$, $\sum_j \mathbb{P}(\xi_j^n(B) > 1) \underset{n\to\infty}{\longrightarrow} 0$.*

We deduce from this theorem a corollary for simple point processes, which we later make an extend use of.

Corollary 0.4. *Consider (ξ_j^n) a null array of Dirac measures on $\mathbb{R}^d$, and ξ a Poisson point measure on $\mathbb{R}^d$ with intensity $m \in \mathcal{M}(\mathbb{R}^d)$. Set for any $n \in \mathbb{N}$, $\xi^n = \sum_j \xi_j^n$. Assume that the sequence of their intensity measures $(\mathbb{E}(\xi_n))$ converges weakly towards m, i.e. for any $B \in \mathcal{B}(\mathbb{R}^d)$ such that $m(\partial B) = 0$, $\mathbb{E}(\xi^n(B))$ converges towards $m(B)$. Then ξ^n converges in distribution towards ξ.*

Proof :
First the random variables ξ_j^n are Dirac masses and have thus values in $\{0,1\}$, so that condition (ii) in Theorem 0.3 is satisfied. Besides, for any relatively compact set $B \in \mathcal{B}(\mathbb{R}^d)$ satisfying $m(\partial B) = 0$,

$$\sum_j \mathbb{P}(\xi_j^n(B) > 0) = \sum_j \mathbb{P}(\xi_j^n(B) = 1) = \mathbb{E}(\xi^n(B)),$$

and then condition (i) in Theorem 0.3 is implied by the weak convergence of the intensity measures $\mathbb{E}(\xi^n)$ towards m. □

0.1.3 Spectrally positive Lévy processes

This section is composed of results that can mostly be found in [Ber96] or [Kyp06]. It consists in a summary of the main points concerning spectrally positive Lévy processes, and emphasizes properties that will later be useful for the study of splitting trees and their genealogies.

We consider a real-valued Lévy process $X = (X_t)_{t\geq 0}$ (that is, X is a càd-làg process with independent and stationary increments), which we will suppose spectrally positive, meaning that it has no negative jumps. We assume furthermore in this section that X is starting at 0 a.s., and denote by $\mathbb{P}$ its law. This Lévy process is characterized by its Laplace exponent ψ defined for all $\lambda \geq 0$ by

$$\mathbb{E}\left(e^{-\lambda X_t}\right) = e^{t\psi(\lambda)},$$

and the Lévy-Khintchine formula gives :

$$\psi(\lambda) := d\lambda + \frac{b^2}{2}\lambda^2 - \int_{(0,\infty)} (1 - e^{-\lambda r} - \lambda h(r))\Lambda(\mathrm{d}r), \tag{1}$$

where h is some arbitrary truncation function on $\mathbb{R}$ (a truncation function h on $\mathbb{R}^d$ is a continuous bounded function from $\mathbb{R}^d$ to $\mathbb{R}^d$ satisfying $h(x) = x$ in a neighborhood of 0). The Lévy measure Λ is a measure on $(\mathbb{R}_+^*, \mathcal{B}(\mathbb{R}_+^*))$ satisfying $\int (1 \wedge |u|^2)\Lambda(\mathrm{d}u) < \infty$. The coefficient b is named Gaussian coefficient, and the coefficient d depends on the choice of the truncation function.

The paths of X have **finite variation** (on every compact time interval) a.s. iff $b = 0$ and $\int (1 \wedge |r|)\Lambda(\mathrm{d}r) < \infty$. In this case, the integral $\int_{(0,\infty)} h(r)\Lambda(\mathrm{d}r)$ is finite a.s., and we can reexpress the Laplace exponent as

$$\psi(\lambda) := -d'\lambda - \int_{(0,\infty)} (1 - e^{-\lambda r})\Lambda(\mathrm{d}r), \tag{2}$$

where d' is called the drift coefficient and characterizes X together with the Lévy measure Λ. It is in particular the case if X is a **subordinator**, i.e. if X has increasing paths a.s., and d' is then nonnegative.

In the sequel, we will sometimes deal with **killed subordinators** : by killed subordinator at a random time T we mean that the value of the process at any time $t \geq T$ is replaced by $+\infty$. By subordinator killed at rate k we mean a killed subordinator at an independent exponential time with parameter k.

Consider the case where X is not a subordinator. The Laplace exponent ψ is infinitely differentiable, strictly convex, and satisfies $\psi(0) = 0$ and $\lim_{\lambda\to\infty} \psi(\lambda) = +\infty$. In particular, $\psi'(0^+) = -\mathbb{E}(X_1) \in [-\infty, +\infty)$. Thus ψ has at most one root besides 0. We denote by η the largest one, and $\eta = 0$ if and only if $\psi'(0^+) \geq 0$. Moreover, X drifts to $+\infty$ (resp. oscillates, drifts to $-\infty$) if and only if $\psi'(0^+)$ is negative (resp. zero, positive). Then we say that X is respectively supercritical, critical or subcritical. Note that if X is supercritical, $\eta > 0$, and that otherwise $\eta = 0$.

Furthermore, the function ψ is a bijection from $[\eta, \infty)$ to $\mathbb{R}_+$ and we define its inverse $\phi : \mathbb{R}_+ \to [\eta, \infty)$. Denote by T^A the **first entrance time** of X in the Borel set A, that is

$$T^A := \inf\{t \geq 0, X_t \in A\},$$

and write T^x for $T^{\{x\}}$. Then we have the following result.

Proposition 0.5. *The process $(T^{-x})_{x\geq 0}$ is a subordinator with Laplace exponent $-\phi$ killed at rate η. In particular,*

$$\mathbb{P}(T^{-x} < \infty) = e^{-\eta x}.$$

Note that for all $x > 0$, since X has no negative jumps, X is a.s. continuous at T^{-x}, and $T^{-x} = T^{(-\infty,-x)}$ a.s.

Finally we introduce the **scale function**, which is in particular useful for solving exit problems (see e.g. [Kyp06, Chapter 8]) : W is defined as the unique increasing continuous function from $\mathbb{R}_+$ to $\mathbb{R}_+$ with Laplace transform

$$\int_{(0,\infty)} e^{-\lambda x} W(x)\mathrm{d}x = \frac{1}{\psi(\lambda)}, \quad \lambda > \eta. \tag{3}$$

In the infinite variation case, W is differentiable on $\mathbb{R}_+^*$ with continuous derivative (see remark after Lemma 8.2, and Exercise 8.4 in [Kyp06]). In the finite variation case, W has left and right derivatives on $\mathbb{R}_+^*$. Moreover, according to [Kyp06, Lemma 8.6], when X is not a subordinator, $W(0)$ is equal to $-1/d'$ (where $d' < 0$ is the drift) in the finite variation case, and is zero in the infinite variation case. Finally, we have the following property :

Proposition 0.6. *For all $a, b > 0$,*

$$\mathbb{P}(T^{-a} < T^{(b,\infty)}) = \frac{W(b)}{W(a+b)}.$$

Local time and excursions

Let X be a spectrally positive Lévy process with Laplace exponent ψ given by formula (1), and denote by $(\mathcal{F}_t)$ the natural filtration associated with X, i.e. for all $t \geq 0$,

$$\mathcal{F}_t = \sigma\{X_s,\ s \leq t\}.$$

We define its past supremum $\bar{X}_t := \sup_{[0,t]} X$ for all $t \geq 0$. Then the reflected process $X - \bar{X}$ is a Markov process in the filtration $(\mathcal{F}_t)$ (and also in its own natural filtration), for which one can construct a local time at 0 and develop an excursion theory. For more details about the following results, see chapter IV in [Ber96].

Local times For the construction of a local time at 0 for $X - \bar{X}$ (which we will also name local time at the supremum for X), we have to distinguish the case of infinite variation, where 0 is regular for X w.r.t. the open half-line $(0, \infty)$, from the case of finite variation, where 0 is irregular w.r.t. the open half-line $(0, \infty)$.

According to Theorem IV.4 in [Ber96], when X has infinite variation, we denote by L a local time at 0 for $X - \bar{X}$, and the mapping $t \mapsto L(t)$ is non decreasing and continuous. Any other local time at 0 for $X - \bar{X}$ differs then from L in a positive multiplicative constant. If X has finite variation and if 0 is irregular for X, we set

$$L(t) := \sum_{i=0}^{\imath(t)} \tau_i,$$

where $\imath(t)$ represents the number of jumps of the supremum up until time t - i.e. the number of zeros of the reflected process up until time t, and $(\tau_i)_{i \geq 0}$ is a sequence of i.i.d. random exponential variables with arbitrary parameter, independent from X. Then L is a local time at the supremum for X, but is only right-continuous. Moreover, L is not adapted to the filtration $(\mathcal{F}_t)$, and to make up for that problem we replace $(\mathcal{F}_t)$ by $(\mathcal{G}_t) := (\mathcal{F}_t \vee \sigma(L_s,\, s \leq t))$.

We can then define in both cases the inverse of L : for all $t \geq 0$, set

$$L^{-1}(t) := \inf\{s \geq 0,\ L(s) > t\}.$$

The process L^{-1} is a killed subordinator, and is adapted to $(\mathcal{G}_{L^{-1}(t)})$. In the infinite variation case, for all $t \geq 0$, $L(L^{-1}(t)) = t$ (while this is false when X has finite variation). On the other hand, in both cases we have

$$L^{-1}(L(t)) = \inf\{L^{-1}(u),\ L^{-1}(u) > t\} = \inf\{s > t,\ \bar{X}_s - X_s = 0\}.$$

Finally we define for all $t > 0$

$$L^{-1}(t-) := \lim_{s \to t-} L^{-1}(s) = \inf\{s \geq 0,\ L(s) \geq t\}.$$

Let us now briefly introduce the so-called *(ascending)* ***ladder height process*** H^+ of the Lévy process X. This process is a time-change of the supremum process of X defined as follows :

$$H^+ = \bar{X} \circ L^{-1},$$

where we recall that for any $t \geq 0$, $\bar{X}(t) = \sup_{0 \leq s \leq t} X(s)$. In particular, the ladder height process is a (possibly killed) subordinator (see [Ber96, Ch.VI]).

Excursion theory We denote by $\mathscr{E}$ the set of excursions of $X - \bar{X}$ away from 0 : $\mathscr{E}$ is the set of the càd-làg functions ϵ with no negative jumps for which there exists $\zeta = \zeta(\epsilon) \in (0, \infty]$, which will be called the lifetime of the excursion, and such that $\epsilon(0) = 0$, $\epsilon(t)$ has values in $(-\infty, 0)$ for $t \in (0, \zeta)$ and in the case where $\zeta < \infty$, $\epsilon(\zeta) \in [0, \infty)$. We endow $\mathscr{E}$ with the topology induced by the Skorokhod topology.

We consider the process $e = (e_t)_{t \geq 0}$ with values in $\mathscr{E} \cup \{\partial\}$ (where ∂ is an additional isolated point), defined by :

$$e_t := \begin{cases} ((X - \bar{X})_{s + L^{-1}(t-)}, 0 \leq s < L^{-1}(t) - L^{-1}(t-)) & \text{if } L^{-1}(t-) < L^{-1}(t) \\ \partial & \text{else} \end{cases}$$

Then according to Theorem IV.10 in [Ber96], if X does not drift to $-\infty$, then 0 is recurrent for the reflected process, and $(t, e_t)_{t\geq 0}$ is a Poisson point process with intensity $c\ \mathrm{d}t\ N(\mathrm{d}\epsilon)$, where c is some constant depending on the choice of L, and N is a measure on $\mathscr{E}$. Else, $(t, e_t)_{t\geq 0}$ is a Poisson point process with intensity $c\ \mathrm{d}t\ N(\mathrm{d}\epsilon)$, stopped at the first excursion with infinite lifetime.

In the same way, we denote by $\mathscr{E}'$ the set of excursions of X away from 0 : $\mathscr{E}'$ is the set of the càd-làg functions ϵ with no negative jumps for which there exists $\zeta = \zeta(\epsilon) \in (0, \infty]$, and such that $\epsilon(0) = 0$, $\epsilon(t)$ has values in $\mathbb{R}^*$ for $t \in (0, \zeta)$, and $\epsilon(\zeta) = 0$ if $\zeta < \infty$. We then introduce $\chi(\epsilon) := \inf\{t \in (0, \zeta],\ \epsilon(t) \in [0, \infty)\}$. We endow $\mathscr{E}'$ with the topology induced by the Skorokhod topology.
Denoting by $\mathscr{L}$ a local time at 0 of X and by $\mathscr{L}^{-1}$ its inverse, we define the process $e' = (e'_t)_{t\geq 0}$ with values in $\mathscr{E}' \cup \{\partial\}$

$$e'_t := \begin{cases} (X_{s+\mathscr{L}^{-1}(t-)}, 0 \leq s < \mathscr{L}^{-1}(t) - \mathscr{L}^{-1}(t-)) & \text{if } \mathscr{L}^{-1}(t-) < \mathscr{L}^{-1}(t) \\ \partial & \text{else} \end{cases}$$

If X has no Gaussian component, any excursion $e'_t \in \mathscr{E}'$ first visits $(-\infty, 0)$, and we necessarily have $\chi(e'_t) > 0$ (but possibly infinite). On the other hand, if X has a Gaussian component, it can creep upwards and then $\chi(e'_t) \in [0, \infty]$. Again, according to Theorem IV.10 in [Ber96], e' is a Poisson point process with intensity $c'\ \mathrm{d}t\ N'(\mathrm{d}\epsilon)$, stopped if X is subcritical at the first excursion with infinite lifetime. Here c' is some constant depending on the choice of $\mathscr{L}$ and N' a measure on $\mathscr{E}'$.

Finally, we describe some marginals of N and N' in the proposition below, for which we refer to [Kyp06, Th.6.15 and (8.29)], [Ber91, (3)] and [Ber92, Cor.1].

Proposition 0.7. *We have for all $z, x > 0$:*

(i) *If X has finite variation,*

$$N(-\epsilon(\zeta-) \in \mathrm{d}x,\ \epsilon(\zeta) \in \mathrm{d}z,\ \zeta < \infty) = W(0)e^{-\eta x}\mathrm{d}x\Lambda(x + \mathrm{d}z)$$

(ii) *If X has infinite variation and no Gaussian component (i.e. $b = 0$),*

$$N(-\epsilon(\zeta-) \in \mathrm{d}x,\ \epsilon(\zeta) \in \mathrm{d}z,\ \zeta < \infty) = e^{-\eta x}\mathrm{d}x\Lambda(x + \mathrm{d}z).$$

Moreover, in both cases, under $N(\cdot \mid -\epsilon(\zeta-) = x,\ \zeta < \infty)$, the reversed excursion

$$\big(-\epsilon((\zeta - t)-),\ 0 \leq t < \zeta\big)$$

is equal in law to $(X_t,\ 0 \leq t < T^0)$ under $\mathbb{P}_x(\cdot \mid T^0 < \infty)$.

Finally, the same statement holds replacing N by N' and ζ by χ.

Convergence of Lévy processes

We recall a restricted version of Corollary 3.6 from [JS87, VII.3], that characterizes the convergence of a sequence of Lévy processes.

Proposition 0.8. *Let X_n, X be spectrally positive Lévy processes with respective Laplace exponents*

$$\psi_n(\lambda) := c_n\lambda + \frac{b_n^2}{2}\lambda^2 - \int (1 - e^{-\lambda u} - \lambda h(u))\Lambda_n(\mathrm{d}u)$$

$$\psi(\lambda) := c\lambda + \frac{b^2}{2}\lambda^2 - \int (1 - e^{-\lambda u} - \lambda h(u))\Lambda(\mathrm{d}u)$$

for some common truncation function h. Then $X_n \Rightarrow X$ in $\mathbb{D}(\mathbb{R}_+)$ iff as $n \to \infty$:

(i) $c_n \to c$,

(ii) $b_n^2 + \int h^2 \mathrm{d}\Lambda_n \to b^2 + \int h^2 \mathrm{d}\Lambda$,

(iii) *For any continuous bounded function g satisfying $g(u) = o(|u|^2)$ when $|u| \to 0$ (or equivalently, vanishing on a neighborhood of 0), $\int g\mathrm{d}\Lambda_n \to \int g\mathrm{d}\Lambda$.*

Remark 0.9. *An analogous version of this statement is available for Lévy processes with values in $\mathbb{R}^d$, for which each coordinate is itself spectrally positive. Note in particular that condition* (ii) *is then : for all $1 \leq i, j \leq d$, $b_{n,i,j}^2 + \int h_i h_j \mathrm{d}\Lambda_n \to b^2 + \int h_i h_j \mathrm{d}\Lambda$, where $b_{n,i,j}$ is the element in the matrix b_n with indices (i, j), and h_i denotes the i-th coordinate of h.*

0.1.4 Splitting trees

Definitions

A splitting tree [Gei96, GK97, Lam10] is a random tree where individuals behave independently from one another, have i.i.d. life durations, and give birth at constant rate during their lives. We give here an intuitive description of splitting trees, and we refer to [Lam10, Sec. 4.1] for a formal definition.

A splitting tree is characterized by a σ-finite measure Λ on $(0, \infty)$, called ***lifespan measure***, satisfying

$$\int_{(0,\infty)} (1 \wedge r)\Lambda(\mathrm{d}r) < \infty.$$

The measure Λ characterizes both the birth rate and the distribution of life spans of individuals as follows : Conditional on the lifetime $\chi \in \mathbb{R}_+^*$ of an individual, the birth times and life spans of her offspring are distributed according to a Poisson random measure on $(0, \chi) \times \mathbb{R}_+^*$ with intensity Leb$\otimes\Lambda$. In particular, the number of children of a given individual is possibly infinite (case $\Lambda((0, \infty)) = \infty$). Conversely, in the case of a finite measure Λ with mass b, births arrive at rate b, and the common distribution of the lifespan of individuals is given by the probability measure Λ/b. Note that the birth-death model with birth rate b and death rate d is recovered when taking $\Lambda(\mathrm{d}r) = bde^{-dr}\mathrm{d}r$.

Splitting trees are a generalization of birth-death processes, in the sense that births arrive at constant rate, but life durations do not necessarily follow an exponential distribution. The **width process** Ξ of a splitting tree, which counts the number of individuals alive in the tree over time, is a branching process that is not Markovian, unless the lifespan measure is exponential (Ξ is then a birth-death process) or a Dirac mass at $\{\infty\}$ (Ξ is then a pure birth process or *Yule process*). When Λ is finite, Ξ is called *binary homogeneous Crump-Mode-Jagers process* [CM68, Jag69].

Planar embedding of a splitting tree and labelling of individuals.

We embed splitting trees in a half-plane in such a way that edges are all parallel and that the edges representing the children of a given individual are placed to the right of this individual, from the youngest one (to the left) to the eldest one (to the right) (see Figure 1).
Consider now the (countable) set of individuals alive at a given time τ. In the embedding described above, any individual is placed to the right of her younger siblings, and their descendant, but to the left of her own descendants. Hence this provides a way of **labelling the extant individuals** at time τ. Hereafter we always refer to that order when labelling extant individuals at a given time.

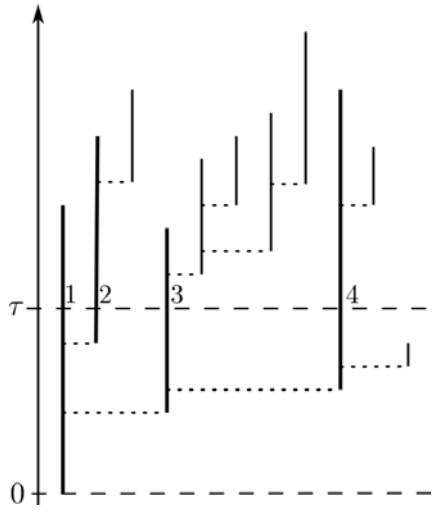

FIGURE 1 – A splitting tree : the vertical axis indicates time ; the horizontal axis has no meaning, but the dotted horizontal lines show filiation. Four individuals are alive at time τ (bold lines), labeled from 1 to 4.

The contour process

Contour techniques are classical in the study of random trees (see e.g. [DL02]). To describe (informally) the contour process of a planar tree, one uses traditionally the picture of a particle moving along edges of the tree. In the « classical » contour process, this particle moves at unit speed, and explores the tree starting from the root and stopping when it first gets back to its starting point. Hence the particle crosses each edge twice, once upwards and once downwards. Imagine now that every upward (continuous) exploration of an edge is replaced by a jump of the particle. This yields a new exploration process, in which the particle visits each point of the tree only once. More precisely, the particle starts at the upward end of the edge corresponding to the ancestor (i.e. the death level of the ancestor), and goes down along this edge at unit speed until it encounters a node (i.e. a birth event). Then the particle jumps to the death level of the newborn and keeps moving this way until it reaches the root of the tree. A graphical representation of the two contour processes described here is given in Figure 2. Note that we used here the term *level* instead of *time*. Hereafter, in the context of splitting trees, we will use this word *level* to denote real time, in which the individuals live, in order to avoid confusion with the index of the contour process, often called *time* itself.

The « new » contour process described above is introduced by A. Lambert in [Lam10] in the setting of splitting trees, and is called *jumping chronological contour process* (or JCCP). We refer to [Lam10] for a formal definition of the JCCP of a splitting tree. The JCCP provides a one-to-one correspondence with the tree, hence many properties of the tree can then be expressed very simply as properties of the JCCP. For example, the number of individuals alive at a given level τ is the number of times that the JCCP hits level τ (see Figure 2). But the main advantage of this process, compared for example to the classical contour process, is that the JCCP of a splitting tree is a Markovian process.

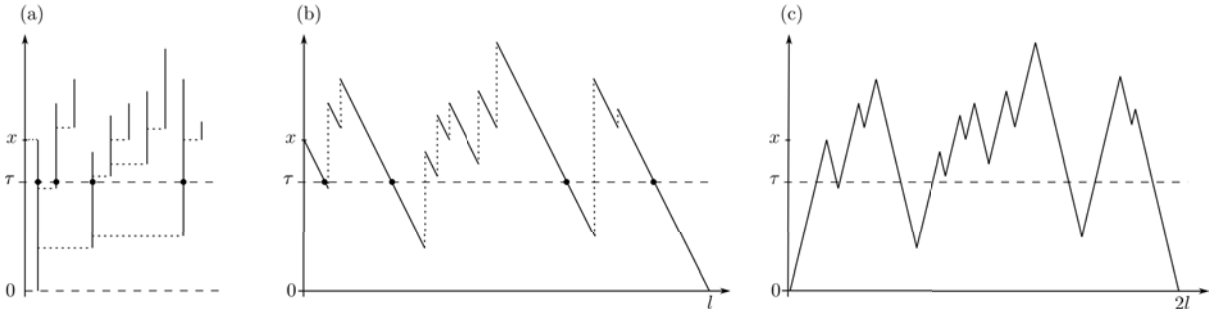

FIGURE 2 – A splitting tree (a) with total length (sum of all edge lengths) l and 4 individuals alive at level τ, its JCCP (b) and its « classical » contour process (c).

Specifically, A. Lambert proves in [Lam10, Th.4.3] that the JCCP of a splitting tree with lifespan measure Λ, truncated up to level τ (i.e. a splitting tree deprived from any of its points with distance to the root greater than τ), is distributed as a spectrally positive Lévy process with finite variation with Laplace exponent

$$\psi(\lambda) := \lambda - \int_{(0,\infty)} (1 - e^{-\lambda r})\Lambda(\mathrm{d}r),$$

reflected below τ and killed upon hitting 0.

This key result, along with the well-developed theory of Lévy processes, enable us to deduce from the JCCP a lot of results on splitting trees. For example, conditional on the lifetime x of the ancestor, and conditional on the size Ξ_τ of the population at a given level τ to be positive, the random variable Ξ_τ follows a geometric distribution with success probability $\mathbb{P}(T^{(0,\infty)} < T^{-\tau})$, which can be expressed in terms of the scale function according to Proposition 0.6. In the next paragraph we show how to study genealogies of splitting trees, using the distribution of the JCCP.

Genealogies of splitting trees

Let us now explain how genealogies of splitting trees are studied with help of the JCCP. Consider τ a fixed positive real number, and $\mathbb{T}$ a splitting tree with lifespan measure Λ. Conditional on the population size Ξ_τ at level τ, denote by H_i, $1 \leq i \leq \Xi_\tau - 1$, the divergence time between individual i and individual $i+1$, i.e. the time elapsed since divergence of individuals i and $i+1$ (we label individuals according to the order described above) : see Figure 3.

Theorem 0.10. *[Lam10, Th.5.4] Conditional on $\Xi_\tau \geq 1$, the sequence of divergence times $(H_i)_{1\leq i\leq \Xi_\tau - 1}$ has the law of a sequence of i.i.d. random variables stopped at its first value greater*

than τ, whose common distribution is that of $\tau - \inf X_t$, where X is a Lévy process with Laplace exponent $\psi(\lambda) := \lambda - \int_{(0,\infty)}(1 - e^{-\lambda r})\Lambda(dr)$, started at τ and killed when exiting $(0, \tau]$.

The sequence $(H_i)_{1\leq i\leq \Xi_\tau - 1}$ characterizes the genealogy of the extant population at level τ : indeed, the time elapsed since divergence of any pair $i, i+k$ of individuals is given by $\max_{i+1\leq j\leq i+k} H_i$ [Lam10, Th.5.4]. In this work, we sometimes use a point measure version $\pi = \sum_{i=1}^{\Xi_\tau - 1} \delta_{(i,H_i)}$ of the sequence $(H_i)_{1\leq i\leq \Xi_\tau - 1}$, yielding a representation of the genealogy that can be seen as the tree spanned by the common ancestors of the n extant individuals of $\mathbb{T}$. Hereafter the sequence $(H_i)_{1\leq i\leq \Xi_\tau - 1}$ and its point measure version are equally called ***coalescent point process*** of the population at τ (see Figure 3).

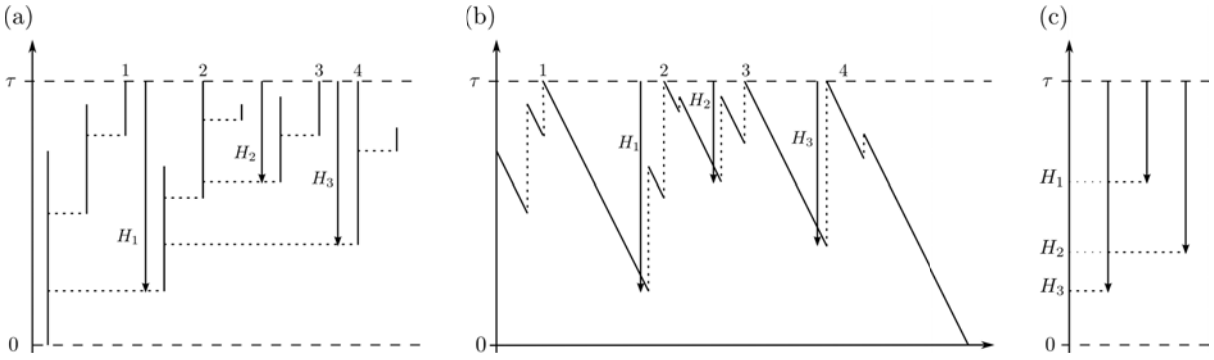

FIGURE 3 – A (truncated) splitting tree (a) with $\Xi_\tau = 4$ individuals alive at level τ, the corresponding contour process (b) and coalescent point process (c). For $i \in \{1, 2, 3\}$, the time H_i elapsed since divergence of individuals i and $i + 1$ is symbolized by a vertical arrow.

Finally, denoting by W the scale function of the Lévy process X introduced in Theorem 0.10, thanks to Proposition 0.6 we have the following result :

Corollary 0.11. *Conditional on $\Xi_\tau \geq 1$, the common distribution of the random variables $(H_i)_{1\leq i\leq \Xi_\tau - 1}$ is given by : for any $x \in [0, \tau]$,*

$$\mathbb{P}(H_1 \leq x \,|\, \Xi_\tau > 1) = \frac{1 - 1/W(x)}{1 - 1/W(\tau)}$$

0.1.5 Population models with mutations

In this thesis, we mainly work with splitting trees enriched with mutations. We specify in this section the framework we adopt concerning mutations, and introduce some tools that are extensively used in the sequel.

Mutation setting

To enrich a population model with mutations, we have to suppose in general that individuals carry types (alleles), or more precisely that each individual is associated with a DNA sequence. We make hereafter two classical assumptions of population genetics concerning mutations :

- **Neutral mutations [Kim84] :** Mutations are supposed to be neutral, meaning that they do not affect the behavior of individuals. Hence enriching the model with mutations does not change the population dynamics.

- **Infinite-site/allele model [Kim69] :** In the infinite-site model, individuals are associated to DNA sequences, and mutations are point substitutions, which are supposed to occur at a site on the sequence that was never hit by a mutation before. In particular, each mutation gives rise to a new allele. Without reference to DNA sequence, this assumption is referred to as the infinite-allele model.

Splitting trees with mutations at birth

In Chapter II, we generalize the notions of splitting tree and JCCP to the framework of population models with mutations. We give here an informal presentation with help of figures, and we refer to Section 0.2.1 and/or Chapter II for formal definitions.
Mutations are seen as marks on the tree, that translate into marks on the contour process. For a splitting tree with finite total length, its so-called *marked JCCP* is then a bivariate Lévy process, which first coordinate is the classical JCCP, and which second coordinate is the counting process of the marks. It provides a one-to-one correspondence with the marked splitting tree, as illustrated in Figure 4.

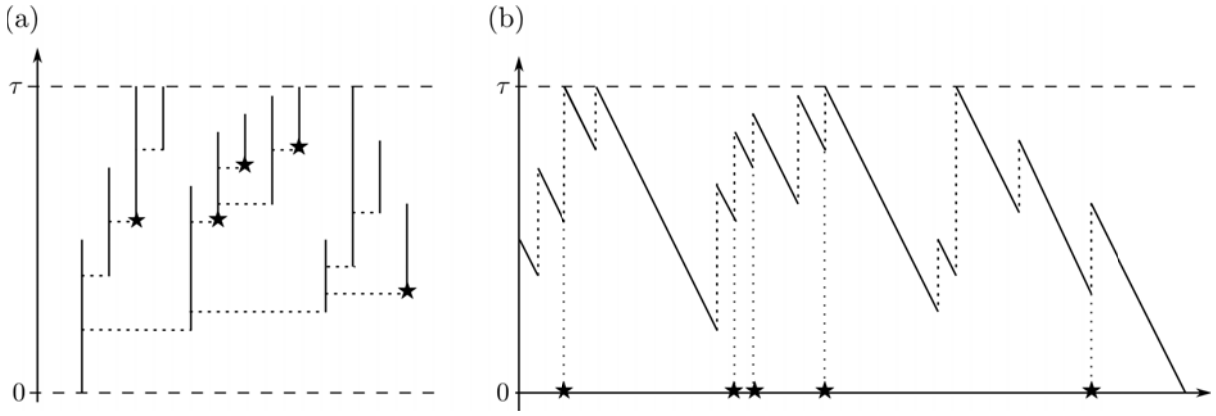

FIGURE 4 – a) A marked splitting tree with mutation events symbolized by stars.
b) The associated marked JCCP, where the counting process of the mutations is not drawn as a jump process on $\mathbb{R}_+$, but is represented by the sequence of its jump times, which are symbolized by stars on the horizontal axis.

Let us now consider genealogies in splitting trees with mutations. The study of mutational patterns leads us to consider, similarly with the definition of the coalescent point process of a tree $\mathbb{T}$ in Section 0.1.4, the tree spanned by both the genealogy of the extant individuals in $\mathbb{T}$ and all the mutation events that affect them. The study of this subtree, later referred to as the *marked coalescent point process*, is made very convenient by the marked JCCP.

We give here an idea of the way genealogies with mutational history can be recovered from the marked JCCP. The following explanations are illustrated in Figure 5, which can be seen as a zoom on a lineage in Figure 4. Consider a splitting tree $\mathbb{T}$ truncated up to a given level τ, and focus on the i-th individual in the extant population at τ. The set of birth times of the ancestors of individual i, up to its coalescence with the rest of the tree, is exactly the set of values taken by the future infimum of the i-th excursion of the JCCP under τ. As a consequence, the mutational history of individual i (up to its coalescence with the rest of the tree) is obtained by selecting the births levels of the ancestors that were hit by a mutation.

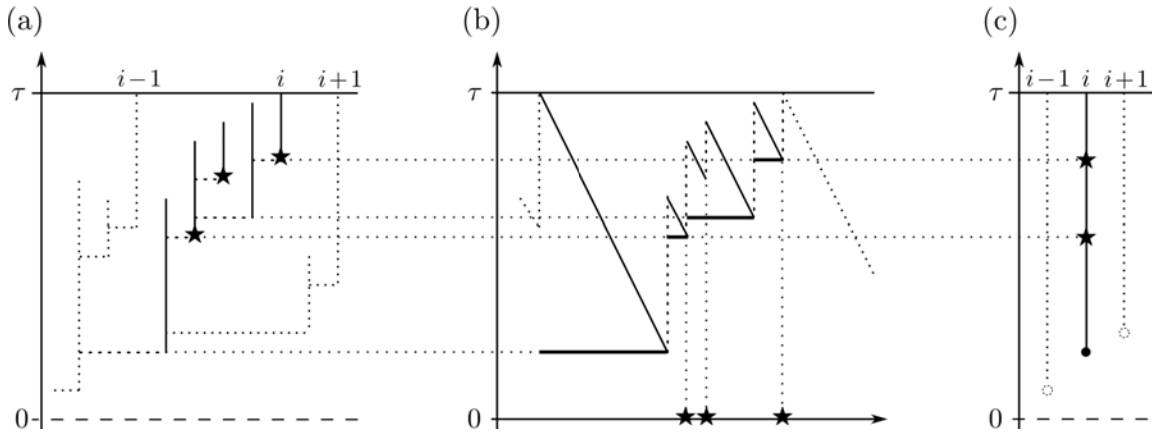

FIGURE 5 – (a) A truncated marked splitting tree, restricted to the extant i-th extant individual, her ancestors (and offspring of ancestors) up to coalescence with the rest of the tree. We also represented (in dotted line) individuals $i-1$ and $i+1$ (and their ancestors).
(b) The corresponding part of the marked JCCP, i.e. (in solid line) its i-th excursion under level τ. In bold line, the future infimum process of this excursion.
(c) The corresponding marked coalescent point process, restricted to the i-th lineage (in solid line).

Further, using the characterization of the (truncated) JCCP as a sequence of excursions of a Lévy process (see Section 0.1.4), along with a time reversal argument (Proposition 0.7), the law of the mutational history of an individual will then be characterized by the image by the past supremum of a Lévy process of a set of marks (that are carried by jumps of this Lévy process). Note that this description is made more rigorous in Section 0.2.1.

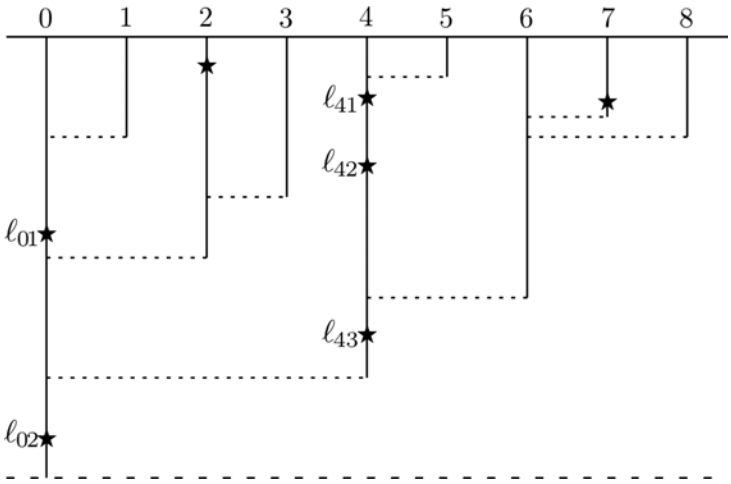

FIGURE 6 – The coalescent point process of a sample of size 9. Mutations are symbolized by stars. Only mutations ℓ_{01}, ℓ_{41} and ℓ_{42} are carried by two individuals, so that here $\xi_2 = 3$. In total, 7 mutations are present on the tree, but mutation ℓ_{02} is carried by every one in the sample, so that $S = 6$.

The frequency spectrum

Finally we introduce the so-called *site frequency spectrum*, that is used in Chapter III to study mutational patterns for samples of individuals associated to DNA sequences. For a sample of size n in the extant population, the site frequency spectrum is the $(n-1)$-tuple $(\xi_1, \ldots, \xi_{n-1})$, where ξ_k is the number of mutations carried by k individuals in the sample : see Figure 6. The sum $S = \xi_1 + \ldots + \xi_{n-1}$ is the total number of *polymorphic sites* (also known as *single nucleotide polymorphisms* in population genetics), i.e. the number of sites at which at least two sequences differ, or equivalently the number of mutations carried by at least one individual, but not all individuals in the sample.

0.2 Outline and statement of results

The main material of this work consists in three articles, each of them accounting for one chapter of the thesis :

- The first two articles [Del13a, Del13b] are to consider as a whole. Their main purpose is to study, in the framework of splitting trees with mutations occurring at birth of individuals, asymptotic properties of genealogies enriched with their mutational history. We study these genealogies with mutations with the help of what we call Lévy processes *with marks*. These processes are to interpret as contour processes of splitting trees, and their *marks* as mutation events. The article [Del13a], published in *Journal of Theoretical Probability*, presents theoretical asymptotic results for Lévy processes with marks. These results are then applied in [Del13b] (submitted to *Stochastic Processes and their Applications)* to establish a limit theorem for the so-called *coalescent point process with marks*, i.e. the genealogy of a splitting tree enriched with its mutational history.
- The third article [ADL14] is joint work with G. Achaz and A. Lambert, and focuses on mutational and genealogical patterns for samples of fixed size in critical branching populations whose scaling limit is a Brownian tree (e.g. critical birth-death trees) (see e.g. [Ald93]), with mutations occurring either at birth or at constant rate during lives of individuals. On the one hand, we provide explicit formulae for the expected site frequency spectrum of the sample. On the other hand, we prove the convergence of the genealogy as the sample size gets large, under various prior distributions on the foundation time of the population. Furthermore, the limiting genealogies with different priors can all be embedded in the same realization of a given Poisson point measure.

In a fourth chapter, we develop some perspectives of the third chapter. We present some preliminary results obtained by applying to real data some theoretical results obtained in [ADL14]. In particular, we use the formula obtained for the site frequency spectrum of our model with fixed time of origin to infer the foundation time of some human subpopulations.

0.2.1 Main results of Chapters I and II

Consider a sequence of splitting trees $(\mathbb{T}_n)$, such that $\mathbb{T}_n$ has lifespan measure Λ_n. Assume that individuals in $\mathbb{T}_n$ carry types, and that each newborn is likely to be hit by a mutation as follows : conditional on her lifetime r, an individual experiences a mutation at her birth with probability $f_n(r)$, where f_n is a continuous function from $\mathbb{R}_+^*$ to $[0,1]$, called *mutation function*. We adopt the framework of neutral mutation in the infinite-allele model, as defined in Section 0.1.5.

The purpose of this work is to provide a limit theorem in a large population asymptotic for the genealogy with mutational history of the sequence $\mathbb{T}_n$. Obtaining such asymptotic results requires to assume on the one hand, the convergence in a certain sense of the population (without mutations) as its size gets large (Assumption A), and on the other hand, about mutations themselves (Assumption B.1 or Assumption B.2).

Before specifying the convergence assumptions, let us introduce some notation. Conditional on the survival of $\mathbb{T}_n$ at a given level, the law of the marked JCCP of $\mathbb{T}_n$ truncated up to this level is characterized by a bivariate Lévy process with finite variation (Z_n, Z_n^{M}), with Lévy measure $\Lambda_n(\mathrm{d}u)\mathbb{B}_{f_n(u)}(\mathrm{d}q)$ and drift $(-1, 0)$, where $\mathbb{B}_p$ denotes the Bernoulli probability measure with parameter p.

Convergence of populations

Assumption A : *There exists a sequence of positive real numbers $(d_n)_{n\geq 1}$ such that as $n \to \infty$, the process defined by*

$$\tilde{Z}_n := \Big(\frac{1}{n} Z_n(d_n t)\Big)_{t\geq 0}$$

converges in distribution to a (necessarily spectrally positive) Lévy process Z with infinite variation. We denote by Λ its Lévy measure and by b its Gaussian coefficient ($b \in \mathbb{R}_+$).

This assumption is to understand as a convergence assumption for the JCCP of the sequence of rescaled populations $(\tilde{\mathbb{T}}_n)$, where $\tilde{\mathbb{T}}_n$ is the splitting tree obtained from $\mathbb{T}_n$ by rescaling the branch lengths by a factor $\frac{1}{n}$.

Convergence of mutations

Two possible hypotheses concerning mutations are considered, both designed to allow the convergence of mutations on the genealogical scale. In the first one, the probability of a child in $\mathbb{T}_n$ to be hit by a mutation is constant, while in the second one, this probability depends on her lifetime.

Assumption B.1 :

(a) *For all $n \geq 1$, for all $u \in \mathbb{R}_+$, $f_n(u) = \theta_n$, where $\theta_n \in [0, 1]$.*

(b) *There exists $\theta \geq 0$ such that $\frac{d_n}{n}\theta_n \to \theta$ as $n \to \infty$.*

Assumption B.2 :

(a) *Their exists $f : \mathbb{R}_+ \to \mathbb{R}_+$ continuous, such that the sequence $\big(u \mapsto \frac{f_n(nu)}{1\wedge u}\big)$ converges uniformly to $u \mapsto \frac{f(u)}{1\wedge u}$ on $\mathbb{R}_+^*$.*

(b) *There exists $\kappa \geq 0$ such that $f(u)/u \to \kappa$ as $u \to 0^+$.*

In Assumption A, it is easy to show that the infinite variation hypothesis for the limiting process Z implies that $d_n/n \to +\infty$ as $n \to \infty$. As a consequence, in Assumption B.1, $\theta_n \to 0$ as $n \to \infty$, corresponding to the classical rare mutation asymptotic.

The asymptotic results we provide in this work concern the limiting genealogy (with mutations) of the rescaled tree $\tilde{\mathbb{T}}_n$, conditioned either on survival at a fixed level τ, or on having $I_n \underset{n\to\infty}{\sim} \frac{d_n}{n}$

extant individuals at τ. We focus here on the second case.

Let us now define the **marked coalescent point process** of $\tilde{\mathbb{T}}_n$. Consider a realization of $\tilde{\mathbb{T}}_n$, and label the I_n individuals alive at τ from 0 to $I_n - 1$ (according to the order defined in Section 0.1.4). Then to the i-th one we associate a simple point measure $\sigma_n^{(i)}$, with values in $(0, \tau) \times \{0, 1\}$, as follows : Consider the lineage of individual i, and assume it contains M mutation events. Denote by m_0 the level where the lineage coalesces with the rest of the tree, and by m_j, $1 \leq j \leq M$ the successive levels (in increasing order) where the mutation events happened (m_1 can coincide with m_0). Then we set

$$\sigma_n^{(i)} := \delta_{(\tau - m_0, 0)} + \sum_{1 \leq j \leq M} \delta_{(\tau - m_j, 1)}.$$

Hence the point measure $\sigma_n^{(i)}$ keeps record of all the mutation events on the i-th lineage, and of the coalescence level of this lineage with the rest of the tree (see Figure 7).

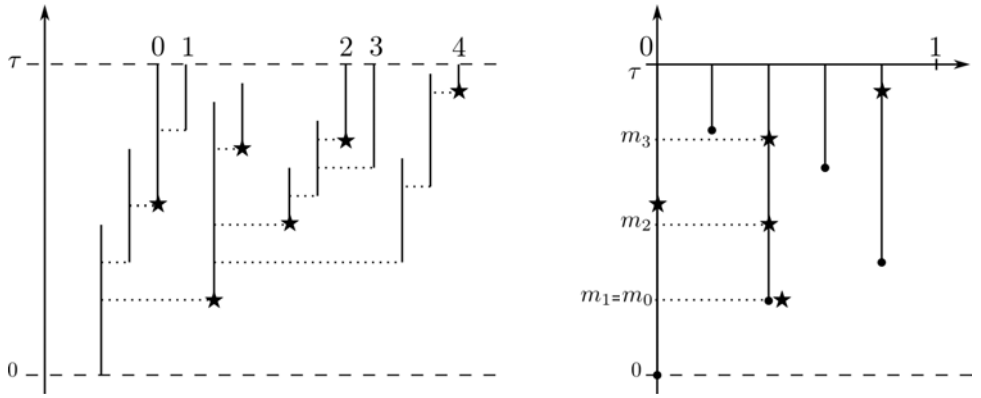

FIGURE 7 – A marked splitting tree truncated up to level τ with 5 extant individuals, and the associated marked coalescent point process (including individual 0, contrary to the definition of Σ_n). The mutation events are symbolized by stars, and dots represent coalescence levels. In this example, the coalescence between the lineages of individuals 1 and 2 coincides with a mutation event, and we have $\sigma_n^{(2)} = \delta_{(\tau - m_0, 0)} + \delta_{(\tau - m_0, 1)} + \delta_{(\tau - m_2, 1)} + \delta_{(\tau - m_3, 1)}$.

Now for all $n \geq 1$, we define the marked coalescent point process Σ_n as a random point measure on $[0, 1] \times \mathcal{M}((0, \tau) \times \{0, 1\})$, such that :

$$\Sigma_n := \sum_{i=1}^{I_n - 1} \delta_{\{\frac{i_n}{d_n}, \sigma_n^{(i)}\}}.$$

The first individual (labeled 0) is on purpose not taken in account.

The main tool used to establish the convergence of the marked genealogical process Σ_n is the marked JCCP of $\tilde{\mathbb{T}}_n$. If we set, according to the rescaling introduced in Assumption A, $\tilde{Z}_n^{\mathrm{M}}(\cdot) := Z_n^{\mathrm{M}}(d_n \cdot)$, the law of the marked JCCP of $\tilde{\mathbb{T}}_n$ truncated up to a given level is then characterized by the process $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$. Chapter I develops technical convergence results for processes linked to the bivariate Lévy processes $((\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}}))_n$, which we now give an overview of.

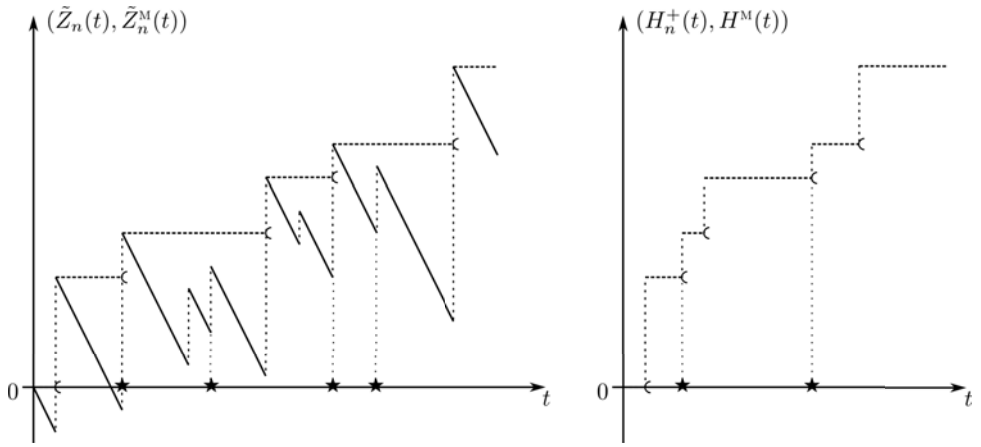

FIGURE 8 – A path of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$, and in dotted line, the past supremum $(\sup_{0\leq s\leq t} \tilde{Z}_n(s))_{t\geq 0}$ of $\tilde{Z}_n$ (left panel). The marked ladder height process $(H_n^+, H_n^{\mathrm{M}})$ of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ (right panel) (Recall that H_n^+ is a time-change of the supremum process of $\tilde{Z}_n$). The subordinators $\tilde{Z}_n^{\mathrm{M}}$ and H_n^{M} are represented by there sequence of jump times, symbolized by stars.

Main results of Chapter I

The main results of Chapter I concern the convergence in distribution of the so called *marked ladder height processes* of the sequence of marked Lévy processes $((\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}}))_n$ under Assumptions A and B.1/B.2. Before giving formal definitions, let us briefly explain how the study of the marked ladder height process is related to the study of genealogies with mutations. Roughly speaking, the marked ladder height process of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ is a bivariate subordinator $(H_n^+, H_n^{\mathrm{M}})$, obtained from the ladder height process H_n^+ of $\tilde{Z}_n$ by putting marks on its jumps, in agreement with the marks on the corresponding jumps of $\tilde{Z}_n$ (see Figure 8). Defining this process enables us to specify the law of the mutational history of an individual. Indeed, the informal characterization given in Section 0.1.5 (see Figure 5), can now be stated as follows : the mutations on a given lineage form an inhomogeneous regenerative set, distributed as the image by H_n^+ of the jump times of H_n^{M}, under the excursion measure of $\tilde{Z}_n$ away from zero.

Note that Assumptions B.1 and B.2 are designed to ensure the convergence in distribution of $(H_n^+, H_n^{\mathrm{M}})$ (and hence of mutations at the genealogical level), but they do not imply, despite Assumption A, the convergence of the process $(\tilde{Z}^n, \tilde{Z}_n^{\mathrm{M}})$: this is even never the case under B.2.

We now define rigorously the marked ladder height process of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$. Denote by $(t, e_{n,t})_{t\geq 0}$ the excursion process of $\tilde{Z}_n$ formed by the excursions from its past supremum, and N_n its excursion measure, as defined in Section 0.1.3. We set for all $t \in [0, L_n(\infty))$

$$\xi_n := \begin{cases} (t, e_{n,t}(\zeta), -e_{n,t}(\zeta-), \Delta\tilde{Z}_n^{\mathrm{M}}(L_n^{-1}(t)))_{t\geq 0} & \text{if } L_n^{-1}(t-) < L_n^{-1}(t) \\ \partial & \text{else} \end{cases},$$

where ∂ is an additional isolated point, L_n is a (suitably normalized) local time at the supremum of $\tilde{Z}_n$ and L_n^{-1} its inverse.
We show that ξ_n is a (possibly killed) Poisson point process whose intensity measure is expressed in terms of the lifespan measure Λ_n and of the mutation function f_n. We then define the process $(H_n^+, H_n^-, H_n^{\mathrm{M}})$ as the trivariate subordinator with no drift and whose jump point process is a.s. equal to the restriction of ξ_n to its last three coordinates, and we call $H_n := (H_n^+, H_n^{\mathrm{M}})$ the

marked ladder height process of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$.

The first important result is the convergence in distribution of the sequence of marked ladder height processes (H_n). Before stating this result, define

$$\mu(\mathrm{d}u, \mathrm{d}q) := \int_0^\infty \mathrm{d}x\ e^{-\eta x}\ \Lambda(x + \mathrm{d}u)\ \mathbb{B}_{f(x+u)}(\mathrm{d}q),$$

and

$$\mu^+(\mathrm{d}u) := \mu(\mathrm{d}u, \{0, 1\}) = \int_0^\infty \mathrm{d}x\ e^{-\eta x}\ \Lambda(x + \mathrm{d}u),$$

where η is the largest positive root of the Laplace exponent of Z. We also denote by W the scale function of Z.

Theorem.
Under ***Assumption B.1****,*

- *if Z does not drift to $-\infty$, the sequence of bivariate subordinators $H_n = (H_n^+, H_n^{\mathrm{M}})$ converges weakly in law to a subordinator $H := (H^+, H^{\mathrm{M}})$, where H^+ and H^{M} are independent, H^+ is a subordinator with drift $\frac{b^2}{2}$ and Lévy measure μ^+, and H^{M} is a Poisson process with parameter θ.*
- *If Z drifts to $-\infty$, the same statement holds but H is killed at rate $k := \frac{1}{W(\infty)}$ and the independence between H^+ and H^{M} holds only conditional on their common lifetime.*

Under ***Assumption B.2****, the sequence of bivariate subordinators $H_n = (H_n^+, H_n^{\mathrm{M}})$ converges weakly in law to a subordinator $H := (H^+, H^{\mathrm{M}})$, which is killed at rate k if Z drifts to $-\infty$. Moreover, H has drift $(\frac{b^2}{2}, 0)$ and Lévy measure*

$$\mu(\mathrm{d}u, \mathrm{d}q) + \rho\delta_0(\mathrm{d}u)\delta_1(\mathrm{d}q),$$

where $\rho := \kappa b^2$.

In particular, under Assumption B.1, we see from the independence of H^+ and H^{M} that the contribution to the mutations in the limit exclusively comes from individuals with vanishing lifetimes.
Under Assumption B.2, if Z has no Gaussian component, the limiting marked ladder height process is a pure jump bivariate subordinator with Lévy measure μ. If Z has a Gaussian component, the fact that the « small jumps » of $\tilde{Z}_n$ generate the Gaussian part in the limit results in a drift for H^+, and possibly additional independent marks that happen with constant rate in time, as under Assumption B.1.
Besides, note that as expected, H^+ is distributed as the (classical) ladder height process of Z.

The second important result of the chapter is an adaptation of a result established in [CD10] to the case of Lévy processes with finite variation. Let L denote a (suitably normalized) local time of Z at its supremum. Before stating this result, note that we have a more general result than the convergence of H_n stated above : we also have the convergence in distribution of the triplet $(H_n^+, H_n^-, H_n^{\mathrm{M}})$ towards a trivariate subordinator whose law is explicitly known.

Theorem. *The following convergence in distribution holds in $\mathbb{D}(\mathbb{R})^4$ as $n \to \infty$:*

$$(\tilde{Z}_n, L_n, H_n^+, H_n^-, H_n^{\mathrm{M}}) \Rightarrow (Z, L, H^+, H^-, H^{\mathrm{M}}),$$

where conditional on (Z, L, H^+, H^-), H^{M} is a Poisson process whose jump process is the jump process of $H^+ + H^-$.

Main results of Chapter II

Chapter II establishes the convergence in distribution, under Assumptions A and B.1/B.2, of the marked coalescent point process Σ_n defined above. For the sake of conciseness, we only state here the results obtained under Assumption B.1.

Recall that L is a local time of the Lévy process Z at its supremum, and H^+ is the ladder height process of Z. Denote by N' the excursion measure of Z away from 0, as defined in Section 0.1.3. Recall that for $\epsilon \in \mathscr{E}'$, $\chi(\epsilon)$ denotes its first entrance time into $[0, \infty)$. Define $\mathscr{E}''$ the set of all càd-làg functions ϵ with lifetime $\zeta < \infty$, such that $\epsilon(0) \geq 0$, $\epsilon(\zeta) = 0$ and $\epsilon(x) > 0$ for all $0 < x < \zeta$. We endow $\mathscr{E}''$ with the topology induced by the Skorokhod topology. Then we define the measure N'' on $\mathscr{E}''$ as the pushforward measure of N' by the mapping

$$\left\{ \begin{array}{ccc} \mathscr{E}' & \longrightarrow & \mathscr{E}'' \\ \epsilon & \longmapsto & (-\epsilon((\chi - t)-))_{0 \leq t < \chi} \end{array} \right. .$$

The following theorem states the convergence of Σ_n, as $n \to \infty$.

Theorem. *Let $(\Theta_i)_{i \geq 0}$ be the sequence of jump times of an independent Poisson process with parameter θ, set $J := \inf\{i \geq 0,\ \Theta_i \geq L(T^0)\}$, and define*

$$\sigma := \delta_{(H^+(L(T^0)-),0)} + \sum_{i=0}^{J-1} \delta_{(H^+(\Theta_i),1)}$$

Then the sequence (Σ_n) converges in distribution towards a Poisson point measure Σ on $[0,1] \times \mathcal{M}((0,\tau) \times \{0,1\})$ with intensity measure $\mathrm{Leb} \otimes \Pi_1$, where Π_1 is a measure on $\mathcal{M}((0,\tau) \times \{0,1\})$ defined by

$$\Pi_1 = N''(\sigma \in \cdot,\ \sup \epsilon < \tau).$$

The fact that the limiting distribution of Σ_n is a Poisson point measure is a consequence of the law of rare events presented above (see Theorem 0.3). Besides, recall that the mutations on a lineage of Σ_n are distributed as the image of H_n^{M} by H_n^+ under the excursion measure of $\tilde{Z}_n$ away from zero. Using the results of Chapter I, it is then not surprising that in the limit, mutations arise as the image of an independent Poisson process by the ladder height process of Z, under a certain excursion measure.

Besides, in the case where the process is **conditioned on survival** at level τ (instead of being conditioned on its population size at τ), the theorem remains valid except for the support of the limiting measure Σ : as a consequence of the geometric distribution of the population size of $\tilde{\mathbb{T}}_n$ at τ, Σ is in this case a Poisson random measure on $[0, \mathrm{e}] \times \mathcal{M}((0,\tau) \times \{0,1\})$, where e is an independent exponential variable with parameter $1/W(\tau)$ (where W is the scale function of the limiting process Z).

A similar theorem is available **under Assumption B.2**, but in particular, the correlation between H^+ and H^{M} makes it slightly trickier to state. Note that as under Assumption B.1, the measure Σ_n still converges towards a Poisson random measures, and mutations on a given lineage arise as the image of H^{M} by H^+, under a certain excursion measure related to the excursion measure of Z away from zero. A significant difference with the case B.1 is that the coalescence time of a lineage might coincide with a mutation event.

Further results are available in the case where Z has **no Gaussian component**. Indeed, Π_1 is then a finite measure, given by

$$\Pi_1 = \int_{(0,\tau)} \mathrm{d}x\, \Lambda((x,\infty))\, \mathbb{P}_x(\sigma \in \cdot,\ T^0 < T^{(\tau,\infty)}).$$

Moreover, the law of σ under $\mathbb{P}_x(\cdot \cap \{T^0 < T^{(\tau,\infty)}\})$ is expressed in terms of the image of an independent Poisson process by an inhomogeneous killed subordinator, whose jump measure is explicitly expressed as a function of Λ and W. Again, a similar result is available under B.2.

Finally, we treat the example where the limiting process Z is the **standard Brownian motion**. Note that this case arises in particular as a scaling limit of the critical birth-death model : if we set for any $n \geq 1$ $\Lambda_n = e^{-r}\mathbb{1}_{r\geq 0}\mathrm{d}r$, and $d_n = \frac{n^2}{2}$, we have the convergence in distribution of $\tilde{Z}_n$ towards the standard Brownian motion.
The distribution of Σ is much simpler to describe here, thanks to the fact that the ladder height process H^+ of the Brownian motion is a deterministic drift. Indeed, the image of an independent Poisson process by H^+ is then a Poisson process itself. Besides, it also implies that H^+ and H^{M} are necessarily independent, so that the theorem stated above under B.1 remains valid, in this particular case, under B.2.
Thus, both under B.1 and B.2, using the properties of the excursion measure of the Brownian motion away from zero, we deduce that the limiting marked coalescent point process is the Poisson point process of the depths of excursions away from zero of the Brownian motion, with depth lower than τ, and with Poissonian mutations on the lineages.

0.2.2 Main results of Chapter III

In this work, we study the genealogy of a sample with fixed size, in a certain class of branching populations with mutations, and aim at obtaining results concerning its allelic partition. To begin with, we prove how different models all result in the same distribution for the genealogy of a sample, namely the law of a critical birth-death model conditioned on its population size, with Poissonian mutations on lineages.

1. A universal law for the genealogy of a sample

Genealogies and sampling in populations conditioned on survival Consider the Brownian case introduced in Section 0.2.1, and define

- a sequence $(\mathbb{T}_N)_{N\in\mathbb{N}}$ of splitting trees, satisfying Assumption A introduced above, with Z the standard Brownian motion,
- a sequence $(f_N)_{N\in\mathbb{N}}$ of continuous functions from $\mathbb{R}_+$ to $[0,1]$ satisfying either Assumption B.1, or Assumption B.2.

As in Section 0.2.1, we assume that any individual in $\mathbb{T}_N$ experiences, conditional on her lifetime r, a mutation at birth with probability $f_N(r)$. Fix now $t > 0$ and assume that $\mathbb{T}_N$ is conditioned on survival at level Nt. We further assume that each individual alive at Nt is independently sampled with probability $p_N \in (0,1)$. Individuals are labeled according to the order defined in Section 0.1.4, and we denote by $I_N = (I_{Nj})_j$ the sequence of indexes of the sampled individuals. Finally, we rescale $\mathbb{T}_N$ by multiplying all its edge lengths by a factor $1/N$. See Figure 9 for a graphical representation of $\mathbb{T}_N$, and of some objects hereafter defined.

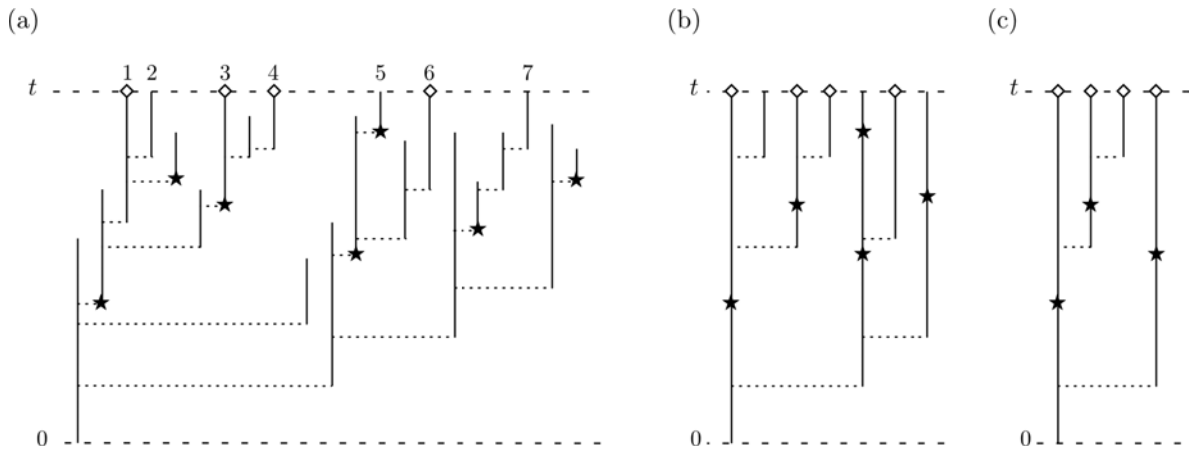

FIGURE 9 – In the three panels (a), (b), (c), the vertical axis indicates time. The horizontal (dotted) lines show filiation. Mutations are symbolized by $\star$ and sampled individuals by $\diamond$.
(a) An example of the (rescaled) tree $\mathbb{T}_N$ with 7 extant individuals at t, where 4 individuals are sampled.
(b) its (marked) coalescent point process (later referred to as Σ_N),
(c) and the (marked) coalescent point process of the sampled individuals.

We are first interested in the distribution of the genealogy of the sampled individuals in $\mathbb{T}_N$, and we consider the model under two slightly different points of view : in case (I), we consider a scaling limit in a large population asymptotic, while in case (II), we consider the example of the critical birth-death process (which satisfies our hypotheses), for which results can be obtained without necessarily having to consider $N \to \infty$. We show here how these two settings lead to the same distribution for the genealogy of a sample, justifying hence the model we later consider for the rest of the paper.

(I) Scaling limit. First, using the results and notation of Chapter II, the (suitably rescaled) marked genealogy Σ_N converges in distribution as $N \to \infty$ towards a Poisson point process on $[0, \mathrm{e}] \times (0, t)$ with intensity $\mathrm{d}l\, x^{-2}\mathrm{d}x$, where e is an independent exponential variable with parameter $1/t$, with independent Poissonian mutations on the lineages. Besides, setting $p_N = p\, d_N/N$ (where d_N is defined in Assumption A and p is a positive real number such that $p_N \in (0, 1)$ for N large enough), the (suitably rescaled) sequence (I_N) of indexes of the sampled individuals (independent of $(\mathbb{T}_N)$), converges towards the sequence of jump times of an independent Poisson process with rate p. The joint convergence of Σ_N with I_N is of course provided by their independence.
As a consequence, from [Lam08] we know that the coalescent point process of the sampled individuals is then distributed as the coalescent point process of a critical birth-death model with rate p conditioned on survival at time t, with independent Poissonian mutations on the lineages.

(II) Critical birth-death tree. Second, let us fix $N \in \mathbb{N}$ and consider the example where $\mathbb{T}_N$ is a critical birth-death tree with rate N conditioned on survival at time t, and assume that (f_N) satisfies Assumption B.1 (constant probability of mutation). Then for any $N \in \mathbb{N}$, the (suitably rescaled) marked coalescent point process Σ_N is distributed as the coalescent point process of a critical birth-death model with rate 1 conditioned on survival at time t, with Poissonian

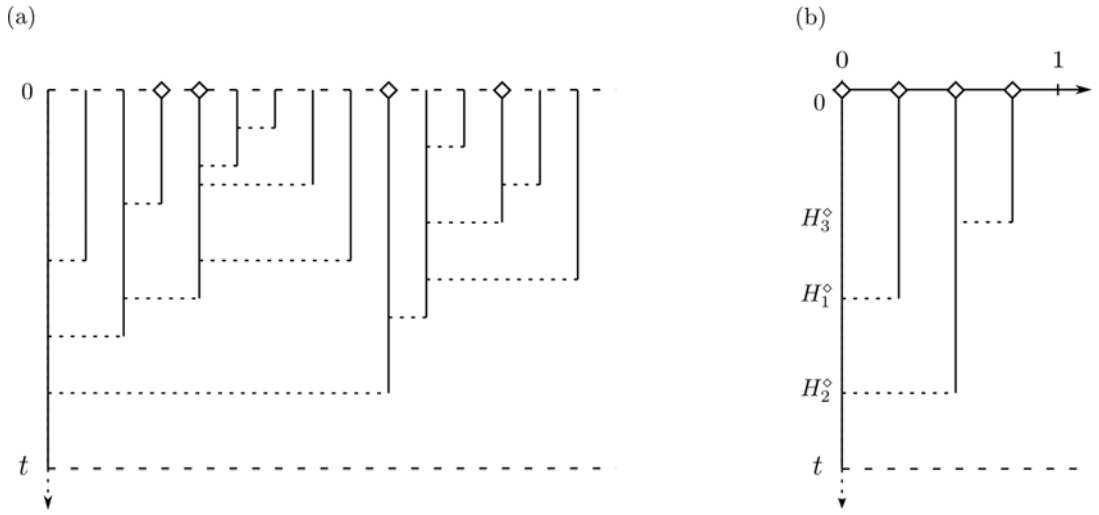

FIGURE 10 – In both figures (a) and (b), the vertical axis indicates time (running backwards).
(a) A graphical representation of the coalescent point process at present time of a (rescaled) tree $\mathbb{T}$ originating at time τ with $n = 4$ sampled individuals (symbolized by $\diamond$). The horizontal lines show filiation.
(b) A graphical representation of the coalescent point process $\pi_n = \sum_{k=1}^{n-1} \delta_{\left(\frac{k}{n}, H_k^\diamond\right)}$ of the sample represented in (a).

mutations on the lineages. Finally, from [Lam08], we get that the coalescent point process of the sampled individuals is then distributed, exactly as above, as the coalescent point process of a critical birth-death model with rate p conditioned on survival at time t, with independent Poissonian mutations on the lineages.

Since the two cases (I) and (II) result in the same distribution for the genealogy of a sample, we limit our study to case (II). Besides, since the mutation schemes arise as independent of genealogies, the next results concerning distributions of genealogies are stated without reference to mutations.

Conditioning on the sample size From now on, consider $\mathbb{T}$ a critical birth-death tree with rate 1. Time is now counted backwards into the past, i.e. « present time » is now time 0, and « u units of time before present » is now time u. Fix $N \in \mathbb{N}$ and $t > 0$, and assume first that $\mathbb{T}$ has been founded Nt units of time ago. As previously, individuals are independently sampled at present time with probability p/N, where $p \in (0, N)$. Besides, we rescale time by a factor $1/N$ (all the edge lengths are then multiplied by a factor $1/N$). We keep the notation $\mathbb{T}$ for the rescaled tree, so that $\mathbb{T}$ is now a critical birth-death tree with rate N, originating at time t.

We now introduce the conditioning on the sample size. We fix $n \in \mathbb{N}$, and we denote by $\mathbb{P}_n^t$ the law of the rescaled tree $\mathbb{T}$ originating at time t and conditioned on its sample size at present time to be equal to n. Note that this conditioning implies *a posteriori* a uniform distribution of the sampled individuals among the total extant population.

The genealogy of the n sampled individuals is characterized by its coalescent point process

$$\pi_n = \sum_{k=1}^{n-1} \delta_{\left(\frac{k}{n}, H_k^\diamond\right)},$$

where for $1 \leq i \leq n-1$, $H_i^\diamond$ is the divergence time between the i-th and the $(i+1)$-th sampled in the rescaled tree $\mathbb{T}$ (see Figure 10). We have the following result :

Theorem. *Under $\mathbb{P}_n^t$, $(H_i^\diamond)_{1\leq i\leq n-1}$ is a sequence of i.i.d. random variables with density function*

$$x \mapsto \frac{p}{(1+px)^2} \frac{1+pt}{pt} \mathbb{1}_{(0,t)}(x).$$

In other words, under $\mathbb{P}_n^t$ the coalescent point process π_n is distributed as the coalescent point process of a critical birth-death model with rate p originating at time t and conditioned on its extant population size to be equal to n.

Remark. *Although we limited here our study to the framework (II) introduced earlier, one could certainly generalize these results (and the upcoming ones) to the scaling limit of case (I). To prove this, one would have to consider a sequence of trees conditioned on their sample size, and then to establish the convergence, in the large population asymptotic, of the marked coalescent point process of the sample. This is however beyond the scope of the present paper.*

Note that we can extend the theorem to the limiting case $t \to \infty$: setting $\mathbb{P}_n^{(\infty)}(\mathbb{T} \in \cdot\,) = \lim_{t\to\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)$, we then have that under $\mathbb{P}_n^{(\infty)}$, $(H_i^\diamond)_{1\leq i\leq n-1}$ is a sequence of i.i.d. random variables with density function $x \mapsto \frac{p}{(1+px)^2} \mathbb{1}_{\mathbb{R}_+}(x)$.

Let us now **randomize the foundation time** of the population, by giving it a (potentially improper) prior distribution of the form $g_i : x \mapsto x^{-i}$. For any $0 \leq i < n$, we denote by $\mathbb{P}_n^{(i)}$ the law of the rescaled tree $\mathbb{T}$, with prior g_i on its time of origin, and conditioned on having n sampled individuals at present time :

$$\mathbb{P}_n^{(i)}(\mathbb{T} \in \cdot,\ T_{\mathrm{or}} \in \mathrm{d}t) = \frac{\mathbb{P}_n^t(\mathbb{T} \in \cdot\,)\mathbb{P}^t(A_n) g_i(t)\,\mathrm{d}t}{\int_0^{+\infty} \mathbb{P}^t(A_n) g_i(t)\,\mathrm{d}t},$$

where A_n denotes the event « the sample size is equal to n », and $\mathbb{P}^t$ is the law of $\mathbb{T}$ originating at t (without conditioning on the sample size). Then we have the following statement.

Proposition. *For any $0 \leq i < n$, the law of $\mathbb{T}$ under $\mathbb{P}_n^{(i)}$ is given by*

$$\mathbb{P}_n^{(i)}(\mathbb{T} \in \cdot\,) = \int_0^{+\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)\, h_n^{(i)}(t)\,\mathrm{d}t,$$

where

$$h_n^{(i)} : t \mapsto pn \binom{n-1}{i} \frac{(pt)^{n-i-1}}{(1+pt)^{n+1}} \mathbb{1}_{\mathbb{R}_+}(t),$$

i.e. the time of origin T_{or} is, under $\mathbb{P}_n^{(i)}$, a random variable, with posterior density function $h_n^{(i)}$.

We then prove that the genealogy of the sample is again distributed as the genealogy of a birth-death process :

Proposition. *Under $\mathbb{P}_n^{(i)}$, the coalescent point process π_n is distributed as the coalescent point process of a critical birth-death model with rate p, with prior g_i on its time of origin and conditioned on its extant population size to be equal to n.*

2. Expected site frequency spectrum

We are now interested in the genetic diversity patterns arising from our model. We provide explicit formulae for the expected site frequency spectrum (see Section 0.1.5) of the sample, under $\mathbb{P}_n^t$, $\mathbb{P}_n^{(0)}$ and $\mathbb{P}_n^{(1)}$. We assume that mutations arise at rate θ on the lineages of the coalescent point process π_n, and we define, for $1 \leq k \leq n-1$, ξ_k the number of mutations carried by k individuals in the sample.

Let us first deal with the case of a uniform prior. In this case, the expected divergence times, under the (suitably rescaled) critical birth-death model with uniform prior on its time of origin and under the Kingman coalescent model, are equal [Ger08]. Now the expected site frequency spectrum can be expressed as a linear combination of the expected divergence times, and besides, the expected site frequency spectrum of the Kingman coalescent is given by $(2\theta/k)_{1\leq k\leq n-1}$ [Wak09]. As a consequence, the expected site frequency spectrum is proportional, under $\mathbb{P}_n^{(0)}$, to that of a sample of the Kingman coalescent :

Proposition. *For any $k \in \{1, \ldots, n-1\}$, $\mathbb{E}_n^{(0)}(\xi_k) = n\theta/pk$.*

The other two cases are proved using a proof method developed in [Lam08]. The formulae we obtain are the following.

Proposition. *For any $k \in \{1, \ldots, n-1\}$, $t \in \mathbb{R}_+^*$, setting $\tau = pt$, we have*

$$\mathbb{E}_n^{\tau}(\xi_k) = \frac{\theta}{p}\left\{\frac{n-3k-1}{k} + \frac{(n-k-1)(k+1)}{k\tau}\right.$$
$$\left. + \frac{(1+\tau)^{k-1}}{\tau^{k+1}}\Big[2\tau^2 - (n-2k-1)2\tau - (n-k-1)(k+1)\Big]\left[\ln(1+\tau) - \sum_{i=1}^{k-1}\frac{1}{i}\left(\frac{\tau}{1+\tau}\right)^i\right]\right\}$$

Proposition. *For any $k \in \{1, \ldots, n-3\}$,*

$$\mathbb{E}_n^{(1)}(\xi_k) = \frac{\theta}{p}\,\frac{n(n-1)}{(n-k)(n-k-2)}\left[\frac{n+k-2}{k} - \frac{2(n-1)}{n-k-1}(\mathcal{H}_{n-1} - \mathcal{H}_k)\right],$$

where for any $k \in \mathbb{N}$, $\mathcal{H}_k = \sum_{j=1}^{k} j^{-1}$.

3. Limit theorem for the coalescent point process as $n \to \infty$

We finally obtain asymptotic results for the law of the genealogy as the sample size gets large. To this aim we let the sampling parameter p depend on n in such a way that $p = n/\alpha$, where α is a fixed positive real number. This choice is natural since the expected size of the sample is then of order n.

Convergence of π_n First define π^t (resp π) as the Poisson point measure on $(0,1)\times(0,\alpha t)$ (resp. $(0,1)\times\mathbb{R}_+^*$) with intensity $\alpha \mathrm{d}l\, x^{-2}\mathrm{d}x\mathbb{1}_{(l,x)\in(0,1)\times(0,\alpha t)}$ (resp. $\alpha \mathrm{d}l\, x^{-2}\mathrm{d}x\mathbb{1}_{(l,x)\in(0,1)\times\mathbb{R}_+^*}$).

Let $(\rho_i)_{i\geq 0}$ be a sequence of i.i.d. exponential random variables with parameter α^{-1}, and define for all $i \geq 0$ the inverse-gamma random variable $\mathrm{e}_i = (\rho_0 + \ldots + \rho_i)^{-1}$. Then for $i \in \mathbb{Z}_+$, define the pair $(\pi^{(i)}, T_{\mathrm{or}}^{(i)})$, where $T_{\mathrm{or}}^{(i)}$ is a positive random variable, and $\pi^{(i)}$ is a Cox process $\pi^{(i)}$, as

$$\mathbb{P}(T_{\mathrm{or}}^{(i)} \in \mathrm{d}t,\ \pi^{(i)} \in \cdot) = \mathbb{P}(\mathrm{e}_i \in \mathrm{d}t)\mathbb{P}(\pi^t \in \cdot).$$

In particular, conditional on $T_{\mathrm{or}}^{(i)} = t$, $\pi^{(i)}$ has the law of the Poisson random measure π^t.

The proof of the next result relies on [Pop04, Th.5], which states the convergence of π_n under $\mathbb{P}_n^t$, towards π^t. Besides, the convergence of π_n under $\mathbb{P}_n^{(0)}$ has already been established in [AP05].

Theorem. *We have the following convergences in distribution as $n \to \infty$:*

a) $$\mathcal{L}(\pi_n,\ \mathbb{P}_n^{(\infty)}) \Rightarrow \pi,$$

b) *and for any $i \geq 0$,* $$\mathcal{L}((\pi_n, \tilde{T}_{\mathrm{or}}),\ \mathbb{P}_n^{(i)}) \Rightarrow (\pi^{(i)}, T_{\mathrm{or}}^{(i)}).$$

We establish next that the random variable $T_{\mathrm{or}}^{(i)}$ is distributed as the i-th atom of the random measure π, where the atoms of π are ordered with respect to their second coordinate. From this observation we derive the following result.

Theorem. *For any $i \in \mathbb{Z}_+$, the measure $\pi^{(i)}$ has the distribution of the random measure obtained from π by removing its $i+1$ largest atoms.*

As a conclusion, in the limit $n \to \infty$, genealogies with different priors on the time of origin can all be embedded in the same realization of the measure π : a realization of the limiting coalescent point process with given prior can be obtained by removing from a realization of π a given number of its largest atoms.

Convergence of the expected frequency spectrum The hypothesis $p = n/\alpha$ yields also the convergence of the expected site frequency spectrum. We obtain from the results of 2. that for any $t \in \mathbb{R}_+^*$ and $i \in \{0, 1\}$, for any $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{E}_n^t(\xi_k) = \alpha\theta/k \quad \text{and} \quad \lim_{n\to\infty} \mathbb{E}_n^{(i)}(\xi_k) = \alpha\theta/k.$$

In other words, under $\mathbb{P}_n^t$, $\mathbb{P}_n^{(0)}$ and $\mathbb{P}_n^{(1)}$, the expected site frequency spectrum of the sample converges, as the size of the sample gets large, towards the expected frequency spectrum of the Kingman coalescent [Wak09, (4.20)].

Chapter I

Lévy processes with marked jumps I : Limit theorems

The article [Del13a] is published in Journal of Theoretical Probability.

1 Introduction

Consider for any $n \in \mathbb{N}$ a σ-finite measure Λ_n on $(0,\infty)$ satisfying $\int (1 \wedge r)\Lambda_n(\mathrm{d}r) < \infty$ and f_n a function from $(0,\infty)$ to $[0,1]$, and assume that the measure $\Lambda_n(\mathrm{d}r)\mathbb{B}_{f_n(r)}(\mathrm{d}q)$, where $\mathbb{B}_p$ denotes the Bernoulli distribution with parameter p, is a Lévy measure. Let $\big((Z_n, Z_n^{\mathrm{M}})\big)_{n\geq 1}$ be a sequence of bivariate Lévy processes with finite variation with values in $\mathbb{R}\times\mathbb{Z}$, such that (Z_n, Z_n^{M}) is characterized by its drift $(-1,0)$ and its Lévy measure $\Lambda_n(\mathrm{d}r)\mathbb{B}_{f_n(r)}(\mathrm{d}q)$. We can interpret this process as a spectrally positive Lévy process with finite variation with additional marks on its jumps ; conditional on the amplitude r of a jump of Z_n, the mark carried by this jump follows a Bernoulli distribution with parameter $f_n(r)$, and Z_n^{M} is then the counting process of these marks.

We consider a rescaled version $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ of (Z_n, Z_n^{M}), and assume the convergence in distribution of the sequence $(\tilde{Z}_n)$, towards a Lévy process Z (with infinite variation, Assumption A). Besides, two different assumptions concerning the marks are considered. In the first one (B.1), (f_n) is a sequence of constant functions vanishing as $n \to \infty$, whereas in the second one (B.2), f_n is a (non constant) function satisfying in particular $f_n(0) = 0$. The goal of this paper is to prove some convergence theorem for the so-called *marked ladder height process* of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$, that we define as a generalization of the classical ladder height process to Lévy processes with marked jumps. These convergence theorems are the first part of a work aiming to obtain asymptotic results for the genealogy of a splitting tree [Gei96, GK97, Lam10] with mutations at birth, enriched with its history of mutations.

Let us explain how these populations can be studied from the marked Lévy processes we just described. First consider a population evolving according to the dynamics of a splitting tree $\mathbb{T}$, that is, a population where individuals give birth at constant rate during their lifetimes to i.i.d. copies of themselves. The jumping chronological contour process (or JCCP) [Lam10] of $\mathbb{T}$ is an exploration process of this tree that provides a one-to-one correspondence with $\mathbb{T}$, and

which distribution is characterized from a spectrally positive Lévy process with finite variation. Assume now that individuals carry types, and that (neutral) mutations may happen at birth of individuals : to each birth event in $\mathbb{T}$ we associate a mark in $\{0,1\}$, which will code for the absence (0) or presence (1) of a mutation. Then the generalization of the JCCP for this splitting tree with marks leads to a characterization of its law by a spectrally positive Lévy process with finite variation, with additional marks on its jumps, as described earlier.

Thus let us interpret our sequence $\big((Z_n, Z_n^{\mathrm{M}})\big)$ as the contour processes of a sequence of marked splitting trees $(\mathbb{T}_n)$. Roughly speaking, the measure Λ_n characterizes the lifetime distribution of the individuals in $\mathbb{T}_n$, and conditional on its lifetime r, an individual has probability $f_n(r)$ to be a mutant. Letting $n \to \infty$, we aim at stating results in a large population asymptotic for $\mathbb{T}_n$, which requires to introduce a rescaling of these populations. Here the convergence assumption on $(\tilde{Z}_n)$ has to be interpreted as the convergence, in a certain sense, of the populations $(\tilde{\mathbb{T}}_n)$ obtained from a proper rescaling of $(\mathbb{T}_n)$.

More precisely, our ultimate goal is to obtain an invariance principle for the genealogy (with mutational history) of the rescaled population $\tilde{\mathbb{T}}_n$, as $n \to \infty$. The characterization of the latter with the help of the JCCP can be obtained from the law of the (marked) future infimum of an excursion of the Lévy process $\tilde{Z}_n$ under a fixed level. By a time reversal argument, this comes to study the (marked) running supremum of $\tilde{Z}_n$ killed upon hitting 0. Here « marking » the future infimum (resp. running supremum) of $\tilde{Z}_n$ means selecting and keeping record of the marks carried by the jumps of the future infimum (resp. running supremum) of $\tilde{Z}_n$. We are thus led to introduce the marked ladder height process of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$: consider H_n^+ the ascending ladder height process of $\tilde{Z}_n$, and put marks on its jumps in agreement with the marks on the corresponding jumps of $\tilde{Z}_n$. Denoting by H_n^{M} the counting process of these marks, the so-called marked ladder height process $(H_n^+, H_n^{\mathrm{M}})$ is then a (possibly killed) bivariate subordinator.

We are here interested in the asymptotic behaviour of these processes under Assumptions A and B.1/B.2 defined above (see also Section 3.1). While Assumption A alone ensures the convergence in distribution of H_n^+ towards the classical ladder height process of Z, Assumptions B.1 and B.2 are designed to allow that of the marked ladder height process. We prove in Section 4 the convergence in law of $(H_n^+, H_n^{\mathrm{M}})$ towards a (possibly killed) bivariate subordinator (H^+, H^{M}), such that H^+ is the ladder height process of Z. Note nevertheless that in this framework there is in general no convergence of the whole mutation process, namely $\tilde{Z}_n^{\mathrm{M}}$. In the case of Assumption B.1, H^+ and H^{M} are independent, and H^{M} is a Poisson process with parameter θ, which arises as the limit of the sequence of constant functions (f_n) after a proper rescaling. This means that the contribution to the marks in the limit exclusively comes from jumps with vanishing amplitudes. This is no longer the case under Assumption B.2, yet additional independent marks can appear if Z has a Gaussian component. In Section 5 we establish the joint convergence in law of $(\tilde{Z}_n, L_n, H_n^+, H_n^{\mathrm{M}})$, where L_n is a local time of $\tilde{Z}_n$ at its supremum. The proof of this result is essentially an adaptation of L. Chaumont and R.A. Doney's paper [CD10], to our specific case of finite variation Lévy processes converging to an infinite variation Lévy process.

2 Preliminaries : Lévy process with marked jumps and marked ladder height process

Let Λ be a measure on $(\mathbb{R}_+^*, \mathcal{B}(\mathbb{R}_+^*))$ satisfying $\int (1 \wedge u)\Lambda(\mathrm{d}u) < \infty$, and f a function from $\mathbb{R}_+^*$ to $[0,1]$. Denote by $\mathbb{B}_r$ the Bernoulli probability measure with parameter r, and consider (X, X^{M}) a bivariate Lévy process with finite variation, with Lévy measure $\Lambda(\mathrm{d}u)\mathbb{B}_{f(u)}(\mathrm{d}q)$ and drift $(-1, 0)$.

These marked Lévy processes will be used in Chapter II to characterize the law of the contour of a splitting tree with mutations at birth, as explained in Section 1. We define now the marked ladder height process of X. This process is a bivariate subordinator, whose first coordinate will be the classical ladder height process of X, and whose second coordinate will keep record of the marks that are present on the jumps of the current supremum of X. It appears naturally in the second paper [Del13b] (Chapter II in this thesis), as a tool to describe the distribution of mutations on the genealogy of a marked splitting tree.

Sticking to the notation introduced in Section 0.1.3 for X and for the local time and excursion process of $X - \bar{X}$, we define for all $t \in [0, L(\infty))$

$$\xi_t := \begin{cases} (t, e_t(\zeta), -e_t(\zeta-), \Delta X^{\mathrm{M}}(L^{-1}(t))) & \text{if } L^{-1}(t-) < L^{-1}(t) \\ \partial & \text{else} \end{cases},$$

where ∂ is an additional isolated point, and $e_t(\zeta)$ (resp. $e_t(\zeta-)$) stands for $e_t(\zeta(e_t))$ (resp. $e_t(\zeta(e_t)-)$).
Here the fourth coordinate $\Delta X^{\mathrm{M}}(L^{-1}(t))$ is 1 or 0 whether or not the jump of X at the right end point of the excursion interval indexed by t carries a mark. Note that the set $\{L^{-1}(t)\}_{t\geq 0}$ of these right end points is exactly the set of record times of X.

Lemma 2.1. *The process ξ is distributed as a Poisson point process on $[0, \mathcal{K}) \times \mathbb{R}_+^* \times \mathbb{R}_+^* \times \{0, 1\}$ with intensity measure*

$$cW(0)\,\mathrm{d}t \cdot \Lambda(x + \mathrm{d}y)\; e^{-\eta x}\mathrm{d}x \cdot \mathbb{B}_{f(x+y)}(\mathrm{d}q),$$

where if X drifts to $-\infty$, $\mathcal{K}$ is an independent exponential variable with parameter $k := c\frac{W(0)}{W(\infty)}$, and else $\mathcal{K} = +\infty$ a.s.

Proof :
We denote by $\tilde{\xi}$ the restriction of ξ to its first three coordinates.
We know from [Ber96, Prop. 0.5.2] and Section 0.1.3 that $\tilde{\xi}$ is distributed as a Poisson point process on $[0, \mathcal{K}) \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ with intensity

$$c\mathrm{d}t\; N(\epsilon(\tilde{\xi}) \in \mathrm{d}y,\, -\epsilon(\tilde{\xi}-) \in \mathrm{d}x),$$

where from Proposition 0.7,

$$N(\epsilon(\tilde{\xi}) \in \mathrm{d}y,\, -\epsilon(\tilde{\xi}-) \in \mathrm{d}x) = W(0)e^{-\eta x}\mathrm{d}x\Lambda(x + \mathrm{d}y),$$

and $\mathcal{K}$ is an independent exponential variable with parameter $cN(\{\epsilon \in \mathscr{E},\; \tilde{\xi}(\epsilon) = \infty\}) = c\frac{W(0)}{W(\infty)}$ if X drifts to $-\infty$, and else $\mathcal{K} = +\infty$ a.s.
Let $B \in \mathcal{B}(\mathbb{R}_+^* \times \mathbb{R}_+^*)$, and $t \geq 0$. Conditional on having an atom of $\tilde{\xi}$ in $[0, t] \times B$, the fourth coordinate of the corresponding atom of ξ follows a Bernoulli distribution with parameter :

$$p(B) := \frac{\int_B f(x + y)\; N(\mathrm{d}y, \mathrm{d}x)}{N(B)}.$$

As a consequence, $\xi([0, t] \times B \times \{1\})$ and $\xi([0, t] \times B \times \{0\})$ follow Poisson distributions with respective parameters $p(B)N(B)ct$ and $(1 - p(B))N(B)ct$, and we deduce that ξ is a Poisson random measure with intensity π, such that for $C \in \mathcal{P}(\{0, 1\})$:

$$\begin{aligned} &\pi([0, t] \times B \times C) \\ &\quad = ct\; N(B)\mathbb{B}_{p(B)}(C) \\ &\quad = ct \int_B \mathbb{B}_{f(x+y)}(C)N(\mathrm{d}y, \mathrm{d}x), \end{aligned}$$

which leads to the result. □

Let $(H^+, H^-, H^{\mathrm{M}})$ be the (possibly killed) trivariate subordinator with no drift and whose jump point process is a.s. equal to the restriction of ξ to its last three coordinates. Here we define H^- only for technical reasons (see Section 5), and hence we now define the marked ladder height process of X as the (possibly killed) bivariate subordinator (H^+, H^{M}). However it will be convenient in the sequel to be also able to name $(H^+, H^-, H^{\mathrm{M}})$; we call it the trivariate ladder height process of X.

Then, as a straightforward consequence of Lemma 2.1 we have

Proposition 2.2. *The marked ladder height process (H^+, H^{M}) is a bivariate subordinator with no drift and Lévy measure*

$$cW(0)\int_0^\infty \mathrm{d}x\; e^{-\eta x}\,\Lambda(x+\mathrm{d}y)\,\mathbb{B}_{f(x+y)}(\mathrm{d}q), \tag{1}$$

and killed at rate $k = c\frac{W(0)}{W(\infty)}$.

Note that H^+ is in fact the ladder height process of X, i.e. for all $t \geq 0$, $H^+(t) = \bar{X}(L^{-1}(t))$ a.s. Moreover, H^{M} is a Poisson process which jumps correspond, in the local time scale, to the marks occurring at record times of X.

3 Definitions and notation

3.1 Convergence assumptions

Let $(\Lambda_n)_{n\geq 1}$ be a sequence of measures on $(\mathbb{R}_+^*, \mathcal{B}(\mathbb{R}_+^*))$ satisfying $\int (1\wedge u)\Lambda_n(\mathrm{d}u) < \infty$ for all n, and $(f_n)_{n\geq 1}$ a sequence of continuous functions from $\mathbb{R}^+$ to $[0,1]$. We consider a sequence of independent bivariate Lévy processes $(Z_n, Z_n^{\mathrm{M}})_{n\geq 1}$ with finite variation, Lévy measure $\Lambda_n(\mathrm{d}u)\mathbb{B}_{f_n(u)}(\mathrm{d}q)$ and drift $(-1, 0)$, where we recall that $\mathbb{B}_r$ denotes the Bernoulli probability measure with parameter r. We first assume

Assumption A : *There exists a sequence of positive real numbers $(d_n)_{n\geq 1}$ such that as $n \to \infty$, the process defined by*

$$\tilde{Z}_n := \Big(\frac{1}{n} Z_n(d_n t)\Big)_{t\geq 0}$$

converges in distribution to a (necessarily spectrally positive) Lévy process Z with infinite variation, and with Lévy measure denoted by Λ.

For all $n \in \mathbb{N}$ and $t \geq 0$, set $\tilde{Z}_n^{\mathrm{M}}(t) := Z_n^{\mathrm{M}}(d_n t)$. In the sequel we always assume that $\tilde{Z}_n(0) = \tilde{Z}_n^{\mathrm{M}}(0) = 0$. With a slight abuse of notation, the law of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ conditional on $(\tilde{Z}_n(0), \tilde{Z}_n^{\mathrm{M}}(0)) = (0,0)$, and the law of Z conditional on $Z(0) = 0$, will both be denoted by $\mathbb{P}$.

Some notation : As in Section 0.1.3, the Laplace exponents ψ_n of Z_n, $\tilde{\psi}_n$ of $\tilde{Z}_n$ and ψ of Z are defined by

$$\mathbb{E}(e^{-\lambda Z_n(t)}) = e^{t\psi_n(\lambda)},\ \mathbb{E}(e^{-\lambda \tilde{Z}_n(t)}) = e^{t\tilde{\psi}_n(\lambda)} \ \text{ and } \ \mathbb{E}(e^{-\lambda Z(t)}) = e^{t\psi(\lambda)}, \quad \lambda \geq 0.$$

We denote by $\tilde{\eta}_n$ (resp. η) the largest root of $\tilde{\psi}_n$ (resp. ψ) and by $\tilde{\phi}_n$ (resp. ϕ) the inverse of $\tilde{\psi}_n$ (resp. ψ) on $[\tilde{\eta}_n, \infty)$ (resp. $[\eta, \infty)$). We denote by $\tilde{W}_n$ (resp. W) the scale function of $\tilde{Z}_n$ (resp. Z). Finally, we denote by $\tilde{\Lambda}_n$ the Lévy measure of $\tilde{Z}_n$.

Remarks about (d_n) **:** Writing for $\lambda \geq 0$, $\mathbb{E}(e^{-\lambda \tilde{Z}_n(t)}) = e^{d_n t \psi_n(\lambda/n)}$, we get from formula (2) that $\tilde{Z}_n$ has drift $-\frac{d_n}{n}$, Lévy measure $\tilde{\Lambda}_n = d_n \Lambda_n(n\cdot)$ and Laplace exponent $\tilde{\psi}_n = d_n \psi(\cdot/n)$. In particular, this gives $\tilde{W}_n(0) = n/d_n$. We state later in Proposition 4.3 that $\tilde{W}_n$ converges pointwise to W as $n \to \infty$, and besides, the assumption of infinite variation of Z ensures $W(0) = 0$. Thereby we know that necessarily $\frac{d_n}{n} \to \infty$ as $n \to \infty$.

Finally, we suggest two possible assumptions for the asymptotic of the marks : in the first one, the probability for a jump of $\tilde{Z}_n$ to carry a mark is constant, while in the second one, this probability is a function of the amplitude of the jump.

Assumption B.1 :

(a) *For all* $n \geq 1$*, for all* $u \in \mathbb{R}_+$*,* $f_n(u) = \theta_n$*, where* $\theta_n \in [0, 1]$*.*

(b) *As* $n \to \infty$*,* $\frac{d_n}{n}\theta_n$ *converges to some finite real number* θ*.*

Assumption B.2 : *There exists* f *a continuous function from* $\mathbb{R}_+$ *to* $\mathbb{R}_+$*, and* $\kappa \in \mathbb{R}_+$*, such that :*

(a) *the sequence* $\left(u \mapsto \frac{f_n(nu)}{1 \wedge u}\right)$ *converges uniformly to* $u \mapsto \frac{f(u)}{1 \wedge u}$ *on* $\mathbb{R}_+^*$*,*

(b) $f(u)/u \to \kappa$ *as* $u \to 0^+$*.*

Note that in B.1, necessarily $\theta_n \to 0$ as $n \to \infty$. Then if we denote by f the limit of the sequence (f_n), we have $f \equiv 0$. Besides, in Assumption B.2 the choice of f_n and f is independent of $\tilde{Z}_n$ and Z.

Remark 3.1. *These two possible assumptions have been chosen so that as* $n \to \infty$*, we have convergence of the set of marks that are carried by jumps of the supremum (which will be reexpressed as sets of mutations on a lineage in Chapter II). However this choice does not imply, despite Assumption A, the convergence of the bivariate process* $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$*. It is even never the case under B.2 : from Proposition 0.8 we see that the convergence as* $n \to \infty$ *of* $\int_{(0,\infty)} f_n(nu)\tilde{\Lambda}_n(\mathrm{d}u)$ *is a necessary condition for that of* $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$*. Now it can be shown that under B.2, this integral behaves as* $n \to \infty$ *like* $\int_{(0,\infty)}(1 \wedge u)\tilde{\Lambda}_n(\mathrm{d}u)$*, which goes to* ∞ *as* $n \to \infty$ *(see Lemma 4.11 for a similar result).*

3.2 Marked ladder height process of $\tilde{Z}_n$

Local times at the supremum

We denote by $\mathcal{F} = (\mathcal{F}_t)_{t \geq 0}$ (resp. $\mathcal{F}_n = (\mathcal{F}_{n,t})_{t \geq 0}$) the natural filtration associated to Z (resp. $\tilde{Z}_n$), that is for all $t \geq 0$,

$$\mathcal{F}_t = \sigma\{Z_s,\ s \leq t\} \text{ (resp. } \mathcal{F}_{n,t} = \sigma\{\tilde{Z}_n(s),\ s \leq t\}).$$

For all $n \geq 1$, let $(\tau_{n,i})_{i \geq 0}$ be a sequence of i.i.d. random exponential variables, independent of $(\tilde{Z}_n)_{n \geq 1}$, with parameter $\alpha_n := \frac{d_n}{n}$. This choice will allow us in the sequel to obtain some convergence properties, in particular for the inverse local time and the ladder height process of $\tilde{Z}_n$. Then, according to Section 0.1.3, we define for $\tilde{Z}_n$ a local time at the supremum as follows :

$$L_n(t) := \sum_{i=0}^{\iota_n(t)} \tau_{n,i},$$

where $\mathfrak{l}_n(t)$ represents the number of jumps of the supremum until time t. We denote by L_n^{-1} the right-continuous inverse of L_n as defined in Section 0.1.3, and replace the filtration $\mathcal{F}_{n,t}$ with $\mathcal{F}_{n,t} \vee \sigma(L_n(s),\, s \leq t)$, so that L_n (resp. L_n^{-1}) is adapted to $(\mathcal{F}_{n,t})$ (resp. to $(\mathcal{F}_{n,L_n^{-1}(t)})$).

As in Section 0.1.3, we introduce the local time at the supremum L for the infinite variation Lévy process Z : we saw that L is defined up to a multiplicative constant, and we require that

$$\mathbb{E}\bigg(\int_{(0,\infty)} e^{-t}\mathrm{d}L_t\bigg) = \phi(1), \tag{2}$$

so that L is uniquely determined. Finally, we denote by L^{-1} its inverse.

Marked ladder height process
For $n \geq 1$, let $(H_n^+, H_n^-, H_n^{\mathrm{M}})$ be the trivariate marked ladder height process of $\tilde{Z}_n$, as defined in section 2. Recall that we are mostly interested in $(H_n^+, H_n^{\mathrm{M}})$ and that we define H_n^- only for technical reasons (see Section 5). For this reason in the sequel, we focus on $(H_n^+, H_n^{\mathrm{M}})$. The results will first be stated in terms of the (bivariate) ladder height process, but their proofs can be easily adapted to the trivariate ladder height process.
Our choice for the normalization of the local times, and the equality $\tilde{W}_n(0) = \frac{n}{d_n}$, along with Proposition 2.2, yields

Proposition 3.2. *The ladder height process $(H_n^+, H_n^{\mathrm{M}})$ is a bivariate subordinator with no drift and Lévy measure*

$$\mu_n(\mathrm{d}y, \mathrm{d}q) := \int_0^\infty \mathrm{d}x\; e^{-\tilde{\eta}_n x}\; \tilde{\Lambda}_n(x + \mathrm{d}y)\; \mathbb{B}_{f_n(n(x+y))}(\mathrm{d}q), \tag{3}$$

and killed at rate $k_n := \frac{1}{\tilde{W}_n(\infty)}$ if $\tilde{Z}_n$ is subcritical.

We also introduce the notation

$$\mu_n^+(\mathrm{d}y) := \mu_n(\mathrm{d}y, \{0,1\}) = \int_0^\infty \mathrm{d}x\; e^{-\tilde{\eta}_n x}\; \tilde{\Lambda}_n(x + \mathrm{d}y) \tag{4}$$

for the Lévy measure of H_n^+. As stated in Section 2, H_n^+ is in fact the ladder height process of $\tilde{Z}_n$, i.e. for all $t \geq 0$, $H_n^+(t) = \bar{\tilde{Z}}_n(L_n^{-1}(t))$ a.s., where $\bar{\tilde{Z}}_n(t)$ denotes the current supremum of $\tilde{Z}_n$ at time t. Moreover, H_n^{M} is a Poisson process with parameter $\lambda_n := \mu_n(\mathbb{R}_+^* \times \{1\})$, so that the random time

$$\mathrm{e}_n := \inf\{t \geq 0,\; H_n^{\mathrm{M}}(t) = 1\} \tag{5}$$

follows on $\{\mathrm{e}_n < L_n(\infty)\}$ an exponential distribution with parameter λ_n.

4 Convergence theorem for the marked ladder height process

4.1 Statement of result

We define

$$\mu(\mathrm{d}u, \mathrm{d}q) := \int_0^\infty \mathrm{d}x\; e^{-\eta x}\; \Lambda(x + \mathrm{d}u)\; \mathbb{B}_{f(x+u)}(\mathrm{d}q),$$

and

$$\mu^+(\mathrm{d}u) := \mu(\mathrm{d}u, \{0,1\}) = \int_0^\infty \mathrm{d}x\; e^{-\eta x}\; \Lambda(x + \mathrm{d}u).$$

Then, we have the following theorem :

Theorem 4.1. *Under Assumption B.1, if Z does not drift to $-\infty$, the sequence of bivariate subordinators $H_n = (H_n^+, H_n^{\mathrm{M}})$ converges weakly in law to a subordinator $H := (H^+, H^{\mathrm{M}})$, where H^+ and H^{M} are independent, H^+ is a subordinator with drift $\frac{b^2}{2}$ and Lévy measure μ^+, and H^{M} is a Poisson process with parameter θ. In the case Z drifts to $-\infty$, the same statement holds but H is killed at rate $k := \frac{1}{W(\infty)}$ and the independence between H^+ and H^{M} holds only conditional on their common lifetime.*

Under Assumption B.2, the sequence of bivariate subordinators $H_n = (H_n^+, H_n^{\mathrm{M}})$ converges weakly in law to a subordinator $H := (H^+, H^{\mathrm{M}})$, which is killed at rate k if Z drifts to $-\infty$. Moreover, H has drift $(\frac{b^2}{2}, 0)$ and Lévy measure

$$\mu(\mathrm{d}u, \mathrm{d}q) + \rho\delta_0(\mathrm{d}u)\delta_1(\mathrm{d}q),$$

where $\rho := \kappa b^2$.

In particular, under Assumption B.2, if Z has no Gaussian component, the limiting marked ladder height process is a pure jump bivariate subordinator with Lévy measure μ. If Z has a Gaussian component, the fact that the « small jumps » of $\tilde{Z}_n$ generate the Gaussian part in the limit results in a drift for H^+, and possibly additional independent marks that happen with constant rate in time, as under Assumption B.1. This rate is proportional to the Gaussian coefficient (provided that $\kappa \neq 0$). Besides, note that as expected, H^+ is distributed as the classical ladder height process of Z. The joint convergence in law of the triplet $(\tilde{Z}_n, H_n^+, H_n^{\mathrm{M}})$ towards (Z, H^+, H^{M}) is established in the next section.

Remark 4.2. *For technical reasons we also need to obtain the convergence in distribution of $(H_n^+, H_n^-, H_n^{\mathrm{M}})$. According to Lemma 2.1, this process is a trivariate pure jump subordinator with Lévy measure*

$$\mathrm{d}x\, e^{-\tilde{\eta}_n x}\, \tilde{\Lambda}_n(x + \mathrm{d}u)\, \mathbb{B}_{f_n(n(x+u))}(\mathrm{d}q),$$

and we can easily adapt the upcoming proofs to get that $(H_n^+, H_n^-, H_n^{\mathrm{M}})$ converges in distribution to a subordinator $(H^+, H^-, H^{\mathrm{M}})$.

4.2 Proof

Consequences of Assumption A

Before proving Theorem 4.1, we state some direct consequences of the convergence of $\tilde{Z}_n$ towards Z. The two following propositions will be frequently used in the sequel and shall be kept in mind by the reader.

Proposition 4.3. (i) *As $n \to \infty$, $\tilde{\phi}_n \to \phi$ uniformly on every compact set of $\mathbb{R}_+$, and in particular $\tilde{\eta}_n \to \eta$.*

(ii) *As $n \to \infty$, $\tilde{W}_n \to W$ uniformly on $\mathbb{R}_+$.*

Proof :

Denote by T_n^x (resp. T^x) the first entrance time of $\tilde{Z}_n$ (resp. Z) in the Borel set $\{x\}$, $x \in \mathbb{R}$. Since Z has no negative jumps it is a.s. continuous at T^{-x}, and we have $\lim_{\varepsilon \to 0+} T^{-(x+\varepsilon)} = T^{-x}$ a.s. Hence as a straightforward consequence of Proposition VI.2.11 in [JS87], we have the convergence in law of T_n^{-x} towards T^{-x}. Now ϕ_n (resp. ϕ) is the Laplace exponent of the process $x \mapsto T_n^{-x}$ (resp. $x \mapsto T^{-x}$) [Ber96, Th.VII.1.1]. The pointwise convergence of $\tilde{\phi}_n$ to ϕ is thus a consequence

of the convergence in distribution of T_n^{-x} towards T^{-x}. The uniform convergence comes from the fact that for all $n \geq 1$, $\tilde{\phi}_n$ is increasing on $\mathbb{R}_+$.
The proof of the pointwise convergence of $\tilde{W}_n$ towards W can be found in [LS12, Prop. 3.1] or can be derived from its definition. Moreover, we have for all $y > x$ $\mathbb{P}(T^{-x} < T^{(y-x,\infty)}) = \frac{W(x)}{W(y)}$ [Ber96, Th.VII.2.8], and then the function $x \mapsto \tilde{W}_n(x)/\tilde{W}_n(y)$ is decreasing. The convergence of $\tilde{W}_n$ towards W is then uniform on every compact set of $\mathbb{R}_+$, and thus uniform on $\mathbb{R}_+$ since the functions are decreasing and bounded from below. □

The Laplace exponent ψ of Z is given for all $\lambda \geq 0$ by :

$$\psi(\lambda) := c\lambda + \frac{1}{2}b^2\lambda^2 - \int (1 - e^{-\lambda u} - \lambda h(u))\Lambda(\mathrm{d}u),$$

where h is a truncation function on $\mathbb{R}$ (see Section 0.1.3). Recall that c depends on the choice of h. Then we have

Proposition 4.4. *Let $(g_n)_{n\geq 0}$ and g be continuous bounded mappings from $\mathbb{R}_+$ to $\mathbb{R}$, where g satisfies $g(u)/u^2 \to K$ as $u \to 0+$ for some constant K. Assume that the mappings $\tilde{g}_n : u \mapsto \frac{g_n(u)}{1\wedge u^2}$ converge uniformly to $\tilde{g} : u \mapsto \frac{g(u)}{1\wedge u^2}$ on $\mathbb{R}_+^*$. Then as $n \to \infty$,*

$$\tilde{\Lambda}_n(g_n) \underset{n\to\infty}{\to} \Lambda(g) + Kb^2.$$

We first prove the following two lemmas. Define $\mathcal{M}_{\mathrm{L}}(\mathbb{R})$ the set of σ-finite measures ν on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying the condition $\int (1 \wedge |u|^2)\nu(\mathrm{d}u) < \infty$.

Lemma 4.5. *Let $(h_n)_{n\geq 0}$ and h be continuous bounded mappings from $\mathbb{R}$ to $\mathbb{R}$, where h satisfies $h(u)/u^2 \to K$ as $u \to 0$ for some constant K. Consider $(\nu_n)_{n\geq 0}$ and ν in $\mathcal{M}_{\mathrm{L}}(\mathbb{R})$ and assume that :*

(i) *There exists $a \in \mathbb{R}$ such that for all continuous bounded function f satisfying $f(u)/u^2 \to K$ as $u \to 0$,*

$$\nu_n(f) \underset{n\to\infty}{\to} \nu(f) + Ka.$$

(ii) *The mappings $\tilde{h}_n : u \mapsto \frac{h_n(u)}{1\wedge u^2}$ converge uniformly to $\tilde{h} : u \mapsto \frac{h(u)}{1\wedge u^2}$ on $\mathbb{R}^*$.*

Then

$$\nu_n(h_n) \underset{n\to\infty}{\to} \nu(h) + Ka.$$

Proof :
First note that since $\nu_n, \nu \in \mathcal{M}_{\mathrm{L}}(\mathbb{R})$, all the integrals considered in the statement of the theorem are finite. Write :

$$\left|\int h_n \mathrm{d}\nu_n - \int h\mathrm{d}\nu - Ka\right| \leq \left|\int (h_n - h)\mathrm{d}\nu_n\right| + \left|\int h\mathrm{d}\nu_n - \int h\mathrm{d}\nu - Ka\right|.$$

The mapping h is continuous and bounded on $\mathbb{R}$, and satisfies $h(u) \sim Ku^2$ when $|u| \to 0$; then (i) implies the convergence to 0 of the term $|\int h\mathrm{d}\nu_n - \int h\mathrm{d}\nu - Ka|$.
Let ε be a positive real number. First observe that (ii) implies that $\tilde{h}_n$ and $\tilde{h}$ can be extended to continuous functions on $\mathbb{R}$ (which we will also denote by $\tilde{h}_n$ and $\tilde{h}$), and we have $\tilde{h}_n(0) \to \tilde{h}(0) = K$. Then (ii) implies for n large enough and any $u \in \mathbb{R}$:

$$|\tilde{h}_n - \tilde{h}|(u) \leq \varepsilon,$$

and then we have $|\int(h_n - h)\mathrm{d}\nu_n| \leq \varepsilon \int(1 \wedge u^2)\nu_n(\mathrm{d}u)$. Now according to (i), the sequence $(\int(1 \wedge u^2)\nu_n(\mathrm{d}u))_n$ converges and is consequently bounded. This proves that $|\int(h_n - h)\mathrm{d}\nu_n|$ tends to 0 and ends the proof. □

Lemma 4.6. *Let g be a continuous bounded function on $\mathbb{R}_+$ such that for some $K \in \mathbb{R}$, $g(u)/u^2 \to K$ as $u \to 0$. Then*

$$\int g\mathrm{d}\tilde{\Lambda}_n \to Kb^2 + \int g\mathrm{d}\Lambda \quad \textit{when } n \to \infty.$$

Proof :
Considering Assumption A, first notice that a straightforward application of Proposition 0.8 yields

(a) For all truncation function h on $\mathbb{R}_+$, $\int h^2\mathrm{d}\tilde{\Lambda}_n \to b^2 + \int h^2\mathrm{d}\Lambda$ as $n \to \infty$.

(b) For any continuous bounded function g such that $g(u) = o(u^2)$ as $u \to 0$, $\int g\mathrm{d}\tilde{\Lambda}_n \to \int g\mathrm{d}\Lambda$.

Then, let h be a truncation function on $\mathbb{R}_+$. Writing $g = Kh^2 + (g - Kh^2)$, we get :

$$\begin{aligned}\left|\int g\mathrm{d}\tilde{\Lambda}_n - (Kb^2 + \int g\mathrm{d}\Lambda)\right| \leq & \left|\int Kh^2\mathrm{d}\tilde{\Lambda}_n - (Kb^2 + \int Kh^2\mathrm{d}\Lambda)\right| \\ & + \left|\int (g - Kh^2)\mathrm{d}\tilde{\Lambda}_n - \int (g - Kh^2)\mathrm{d}\Lambda\right|.\end{aligned}$$

Now since h is a truncation function, thanks to (a) we know that the first term of the right-hand side vanishes as $n \to \infty$. As for the second term, the function $g - Kh^2$ is bounded and satisfies $\lim_{u\to 0} \frac{g(u) - Kh^2(u)}{u^2} = 0$ so that we can apply (b), and $|\int(g - Kh^2)\mathrm{d}\tilde{\Lambda}_n - \int(g - Kh^2)\mathrm{d}\Lambda| \to 0$ as $n \to \infty$. □

Finally, Proposition 4.4 arises as a direct consequence of Lemmas 4.5 and 4.6.

Convergence of the classical ladder process
For all $n \geq 1$ let κ_n be the Laplace exponent of the bivariate ladder process (L_n^{-1}, H_n^+), and denote by κ the Laplace exponent of (L^{-1}, H^+). Note that the condition of normalization (2) imposed to L implies $\kappa(1, 0) = \phi(1)^{-1}$.

Proposition 4.7. *The sequence $(L_n^{-1}, H_n^+)_{n\geq 1}$ converges weakly in distribution to (L^{-1}, H^+).*

Lemma 4.8. *For all $n \geq 1$, $\kappa_n(1, 0) = \tilde{\phi}_n(1)^{-1}$.*

Proof :
Let T_n be the first jump time of the process $\bar{\bar{Z}}_n$. The subordinator L_n^{-1} is a compound Poisson process with rate α_n and jump size distribution $\mathcal{L}(T_n)$ (where $\mathcal{L}(T_n)$ denotes the law of T_n). Therefore we have

$$\begin{aligned}e^{-\kappa_n(1,0)} &= \mathbb{E}(e^{-L_n^{-1}(1)}) \\ &= \sum_{k\geq 0} \frac{(\alpha_n)^k e^{-\alpha_n}}{k!} \mathbb{E}(e^{-T_n})^k \\ &= e^{-\alpha_n(1-\mathbb{E}(e^{-T_n}))}.\end{aligned}$$

Now the variable T_n is a.s. finite and from Theorem 8.1 and Lemma 8.6 in [Kyp06], we get

$$\mathbb{E}(e^{-T_n}) = 1 - \frac{n}{d_n \tilde{\phi}_n(1)}.$$

Since $\alpha_n = \frac{d_n}{n}$, we get $\alpha_n(1 - \mathbb{E}(e^{-T_n})) = \tilde{\phi}_n(1)^{-1}$ and in consequence $\kappa_n(1,0) = \tilde{\phi}_n(1)^{-1}$. □

Proof of Proposition 4.7 :
According to [JS87, Th.VII.3.4], proving the convergence of the Laplace exponents of (L_n^{-1}, H_n^+) is sufficient. Fix $(\alpha, \beta) \in \mathbb{R}_+ \times \mathbb{R}_+$. From Corollary VI.10 in [Ber96], and since $\tilde{Z}_n$ (resp. Z) is not a compound Poisson process (implying its marginal distributions do not have an atom at zero), we know that

$$\kappa_n(\alpha,\beta) = \kappa_n(1,0) \exp\Big\{ \int_0^\infty \mathrm{d}t \int_{(0,\infty)} (e^{-t} - e^{-\alpha t - \beta x}) \frac{1}{t} \mathbb{P}(\tilde{Z}_n(t) \in \mathrm{d}x) \Big\}$$

and

$$\kappa(\alpha,\beta) = \kappa(1,0) \exp\Big\{ \int_0^\infty \mathrm{d}t \int_{(0,\infty)} (e^{-t} - e^{-\alpha t - \beta x}) \frac{1}{t} \mathbb{P}(Z(t) \in \mathrm{d}x) \Big\}.$$

First assume that $\beta = 0$ and $\alpha > 1$. From Assumption A, for all $t > 0$ a.s. the measures $\mathbb{P}(\tilde{Z}_n(t) \in \mathrm{d}x)\mathbb{1}_{x>0}$ converge weakly to $\mathbb{P}(Z(t) \in \mathrm{d}x)\mathbb{1}_{x>0}$. Besides, Lemma 4.8 ensures the convergence of $\kappa_n(1,0) = \tilde{\phi}_n(1)^{-1}$ towards $\phi(1)^{-1} = \kappa(1,0)$ as $n \to \infty$. Then using Fatou's Lemma we obtain

$$\liminf \kappa_n(\alpha, 0) \geq \kappa(\alpha, 0).$$

But from (3), p. 166 in [Ber96], $\kappa_n(\alpha,0)\hat{\kappa}_n(\alpha,0) = \alpha$, so that

$$\liminf \frac{1}{\hat{\kappa}_n(\alpha,0)} \geq \frac{1}{\hat{\kappa}(\alpha,0)}, \text{ and then } \quad \limsup \hat{\kappa}_n(\alpha,0) \leq \hat{\kappa}(\alpha,0),$$

where $\hat{\kappa}_n$ and $\hat{\kappa}$ refer respectively to $\hat{\tilde{Z}}_n = -\tilde{Z}_n$ and $\hat{Z} = -Z$. Then replacing $\tilde{Z}_n$ by $\hat{\tilde{Z}}_n$ in the above arguments entails $\kappa_n(\alpha,0) \to \kappa(\alpha,0)$ as $n \to \infty$. The same arguments hold for $\alpha \in (0,1)$ by exchanging lim sup and lim inf.
Now using the notation of Chapter VI in [Ber96], let τ be an independent exponential variable with parameter $q > 0$, and define $G_\tau^{(n)} := \sup\{t < \tau, \bar{\tilde{Z}}_n(t) = \tilde{Z}_n(t)\}$ (resp. $G_\tau := \sup\{t < \tau, \bar{Z}(t) = Z(t)\}$) the last zero of the reflected process $\bar{\tilde{Z}}_n - \tilde{Z}_n$ (resp. $\bar{Z} - Z$) before τ.

Claim : $(G_\tau^{(n)}, \bar{\tilde{Z}}_n(\tau))$ converges in law towards $(G_\tau, \bar{Z}(\tau))$.

From the weak convergence of $\tilde{Z}_n$ towards Z and using the Skorokhod representation theorem, we can assume that $\tilde{Z}_n \xrightarrow{a.s.} Z$, and it is sufficient to prove the a.s. convergence of $(G_\tau^{(n)}, \bar{\tilde{Z}}_n(\tau))$ towards $(G_\tau, \bar{Z}(\tau))$.

First, the a.s. convergence of $\bar{\tilde{Z}}_n(\tau)$ towards $\bar{Z}(\tau)$ is straightforward from Proposition VI.2.11 in [JS87]. Let us now prove that $G_\tau^{(n)} \xrightarrow{a.s.} G_\tau$.

Using time reversal and considering the infimum process of the reversed reflected process, a direct adaptation of the proofs of Propositions VI.2.4 and VI.2.11 in [JS87] allows us to obtain the inequality $\liminf G_\tau^{(n)} \geq G_\tau$ a.s.

Let us now prove that $\limsup G_\tau^{(n)} \leq G_\tau$ a.s. Consider a realization $(\tilde{z}_n, z, g_T^{(n)}, g_T, T)$ of $(\tilde{Z}_n, Z, G_\tau^{(n)}, G_\tau, \tau)$. Assume there exists $t \in (0,T)$, and a subsequence $(k_n)_{n\in\mathbb{N}}$ satisfying for all n, $g_T^{(k_n)} < t < g_T$.

For all $n \in \mathbb{N}$, by definition of $g_T^{(k_n)}$, on $[g_T^{(k_n)}, T)$ the current supremum $\bar{\tilde{z}}_{k_n}$ of $\tilde{z}_{k_n}$ is equal to a constant $s^{(k_n)}$. Define $\bar{z}$ the current supremum of z and $g'_T := \sup\{u < g_T,\ \bar{z}(u) - z(u) = 0\}$ the penultimate zero of $\bar{z} - z$ before T. On $[g'_T, g_T)$ (resp. on $[g_T, T)$), $\bar{z}$ is equal to a constant s' (resp. s). Applying Proposition VI.2.11 in [JS87] at times t and g_T, we obtain the convergence of $s^{(k_n)}$ towards s and s', which entails $s = s'$. Hence we get the existence of two times $g'_T < g_T$ such that $\bar{z}(g'_T) = z(g'_T) = \bar{z}(g_T) = z(g_T)$. Finally, Proposition VI.4 in [Ber96] shows that such realizations form a negligible set, so that $\limsup G_\tau^{(n)} \leq G_\tau$ a.s. We conclude that $G_\tau^{(n)} \xrightarrow{a.s.} G_\tau$.

The convergence in law of $(G_\tau^{(n)}, \bar{\tilde{Z}}_n(\tau))$ towards $(G_\tau, \bar{Z}(\tau))$ entails, from (1) p.163 in [Ber96], that $\kappa_n(q,0)/\kappa_n(\alpha+q,\beta) \to \kappa(q,0)/\kappa(\alpha+q,\beta)$ as $n \to \infty$. We conclude from the convergence established above. □

Proof of Theorem 4.1
The proof of the theorem will consist in applying Proposition 0.8 to the sequence of bivariate Lévy processes (H_n^+, H_n^{M}). To this aim we first establish the following property.

Proposition 4.9. *The measure $\mu(\,\cdot\,, \{1\})$ is finite, and for any continuous bounded function g on $\mathbb{R}_+$ which is differentiable at 0, we have as $n \to \infty$:*

(i) *Under Assumption B.1,* $\int g(u)\mu_n(\mathrm{d}u, \{1\}) \to \theta g(0)$.
Under Assumption B.2, $\int g(u)\mu_n(\mathrm{d}u, \{1\}) \to \int g(u)\mu(\mathrm{d}u, \{1\}) + \rho g(0)$.
where $\rho = \kappa b^2$ has been defined in Theorem 4.1.

(ii) *Now if $g(0) = 0$,*
Under Assumption B.1, $\int g(u)\mu_n(\mathrm{d}u, \{0\}) \to \int g(u)\mu^+(\mathrm{d}u) + g'(0)\frac{b^2}{2}$.
Under Assumption B.2, $\int g(u)\mu_n(\mathrm{d}u, \{0\}) \to \int g(u)\mu(\mathrm{d}u, \{0\}) + g'(0)\frac{b^2}{2}$.

Furthermore, in both cases, for all $\delta > 0$, the results are still valid if we replace g by $g\,\mathbb{1}_{[0,\delta]}$ or by $g\,\mathbb{1}_{(\delta,\infty)}$.

First of all, to prove this proposition we need the two lemmas below. The first one is deduced from the convergence in law of (H_n^+). The second one is specific to the case B.1.

Lemma 4.10. *Let g be a continuous bounded function from $\mathbb{R}_+$ to $\mathbb{R}$ such that $g(u) = o(u^2)$ as $u \to 0$. We have*

$$\int_{(0,\infty)} \left(\int_0^z g(z-y) e^{-\tilde{\eta}_n y} \mathrm{d}y \right) \tilde{\Lambda}_n(\mathrm{d}z) \underset{n\to\infty}{\to} \int_{(0,\infty)} \left(\int_0^z g(z-y) e^{-\eta y} \mathrm{d}y \right) \Lambda(\mathrm{d}z).$$

Proof :
From Proposition 0.8 and the convergence $H_n^+ \Rightarrow H^+$ established in Proposition 4.7, we get that

$\mu_n^+(g) \to \mu^+(g)$ as $n \to \infty$, where μ^+ denotes the Lévy measure of H^+. Now we deduce from the expression of μ_n^+ given by (4) that

$$\mu_n^+(g) = \int_{(0,\infty)} \tilde{\Lambda}_n(\mathrm{d}z) \int_0^z e^{-\tilde{\eta}_n y} g(z-y)\mathrm{d}y.$$

A similar calculation for the limiting process gives $\mu^+(g) = \int_{(0,\infty)} \Lambda(\mathrm{d}z) \int_0^z e^{-\eta y} g(z-y)\mathrm{d}y$, and the result follows. □

Lemma 4.11. *As $n \to \infty$, we have*

$$\int_{(0,\infty)} \tilde{\Lambda}_n(\mathrm{d}u) \left(\int_0^{1\wedge u} e^{\tilde{\eta}_n(r-u)} \mathrm{d}r \right) \sim \frac{d_n}{n}.$$

Proof :
For all $a > 0$, we have by definition of $\tilde{\phi}_n$ and thanks to formula (2) :

$$\frac{d_n}{n}\tilde{\phi}_n(a) - \int_{(0,\infty)} (1 - e^{-\tilde{\phi}_n(a)u})\tilde{\Lambda}_n(\mathrm{d}u) = a.$$

Then we have

$$1 - \frac{n}{d_n} \int_{(0,\infty)} \frac{1 - e^{-\tilde{\phi}_n(a)u}}{\tilde{\phi}_n(a)} \tilde{\Lambda}_n(\mathrm{d}u) = \frac{n}{d_n} \frac{a}{\tilde{\phi}_n(a)}$$

which leads to

$$\begin{aligned} &\frac{n}{d_n} \int_{(0,\infty)} \tilde{\Lambda}_n(\mathrm{d}u) \left(\int_0^{1\wedge u} e^{\tilde{\eta}_n(r-u)} \mathrm{d}r \right) \\ &\qquad = 1 - \frac{n}{d_n} \frac{a}{\tilde{\phi}_n(a)} - \frac{n}{d_n} \int_{(0,\infty)} \left(\frac{1 - e^{-\tilde{\phi}_n(a)u}}{\tilde{\phi}_n(a)} - \int_0^{1\wedge u} e^{\tilde{\eta}_n(r-u)} \mathrm{d}r \right) \tilde{\Lambda}_n(\mathrm{d}u). \end{aligned}$$

Now it is easy to check that we can apply Proposition 4.4 (further applications of this proposition are detailed in the proof of Proposition 4.9) to get the convergence of

$$\int_{(0,\infty)} \left(\frac{1 - e^{-\tilde{\phi}_n(a)u}}{\tilde{\phi}_n(a)} - \int_0^{1\wedge u} e^{\tilde{\eta}_n(r-u)} \mathrm{d}r \right) \tilde{\Lambda}_n(\mathrm{d}u)$$

towards a finite quantity. Furthermore, we know that $\tilde{\phi}_n(a) \to \phi(a)$ and that $\frac{n}{d_n}$ vanishes as $n \to \infty$, which leads to the announced result. □

Proof of Proposition 4.9 :
We begin with the proof of point (i). Let g be a continuous bounded function on $\mathbb{R}_+$, differentiable at 0. We have :

$$\begin{aligned} \int g(u)\mu_n(\mathrm{d}u, \{1\}) &= \int_{(0,\infty)} \int_0^\infty \mathrm{d}y\ e^{-\tilde{\eta}_n y}\ \tilde{\Lambda}_n(y + \mathrm{d}u)\ f_n(n(u+y))\ g(u) \\ &= \int_{(0,\infty)} \tilde{\Lambda}_n(\mathrm{d}u)\ f_n(nu) \int_0^u \mathrm{d}y\ e^{-\tilde{\eta}_n y}\ g(u-y) \\ &= \int_{(0,\infty)} \tilde{\Lambda}_n(\mathrm{d}u)\ f_n(nu) \int_0^u \mathrm{d}z\ e^{\tilde{\eta}_n(z-u)}\ g(z), \end{aligned}$$

and a similar calculation is available for μ.
Let us first treat the case of Assumption B.2. The calculation above entails

$$\begin{aligned}
&\Big|\int g(u)\mu_n(\mathrm{d}u,\{1\}) - \int g(u)\mu(\mathrm{d}u,\{1\}) - \rho g(0)\Big| \\
\leq\ &\Big|\int \tilde{\Lambda}_n(\mathrm{d}u) f_n(nu)\Big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)\Big) - \int \Lambda(\mathrm{d}u) f(u)\Big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\eta(z-u)} g(z)\Big) - \rho g(0)\Big| \\
&+ \Big|\int \tilde{\Lambda}_n(\mathrm{d}u) f_n(nu)\Big(\int_{1\wedge u}^{u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)\Big) - \int \Lambda(\mathrm{d}u) f(u)\Big(\int_{1\wedge u}^{u} \mathrm{d}z\, e^{\eta(z-u)} g(z)\Big)\Big|. \quad (6)
\end{aligned}$$

First note that the integral $\int_{1\wedge u}^{u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)$ can be rewritten as $\int_0^u \mathrm{d}y\, e^{-\tilde{\eta}_n y} g(u-y)\mathbb{1}_{u-y\geq 1}$. Since the function $z \mapsto g(z)\mathbb{1}_{z\geq 1}$ is bounded and vanishes on $[0,1]$, a simple approximation argument allows us to obtain from Lemma 4.10 the convergence of $\int \tilde{\Lambda}_n(\mathrm{d}u)\big(\int_{1\wedge u}^{u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)\big)$ towards $\int \Lambda(\mathrm{d}u)\big(\int_{1\wedge u}^{u} \mathrm{d}z\, e^{\eta(z-u)} g(z)\big)$. Then, the convergence to 0 of the second term in the right-hand side is obtained using the fact that $|f_n| \leq 1$ for all n, and the uniform convergence on $\mathbb{R}_+$ of $f_n(n\cdot)$ towards f.

Next we focus on the first term. We set $h_n(u) := f_n(nu)\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)$ and $h(u) := f(u)\int_0^{1\wedge u} \mathrm{d}z\, e^{\eta(z-u)} g(z)$. The aim of the next paragraph is to check that the functions h_n and h satisfy the hypotheses of Proposition 4.4, which will entail the convergence to 0 of the first term in the right-hand side of (6).

- The functions $|h_n|$ and $|h|$ can be upper bounded by $\int_0^1 |g(z)|\mathrm{d}z$, which is a finite quantity. Moreover the continuity of g, f_n and f ensures that of h_n and h.

- We have for $u \leq 1$

$$\frac{f(u)}{u} \times \min_{x\in[0,u]}\{g(x)\}\frac{1}{u}\int_0^u e^{-\eta y}\mathrm{d}y \leq \frac{h(u)}{u^2} \leq \frac{f(u)}{u} \times \max_{x\in[0,u]}\{g(x)\}\frac{1}{u}\int_0^u e^{-\eta y}\mathrm{d}y,$$

  Now $\frac{1}{u}\int_0^u e^{-\eta y}\mathrm{d}y \to 1$ as $u \to 0$, and $\lim_{u\to 0}\min_{x\in[0,u]}\{g(x)\} = \lim_{u\to 0}\max_{x\in[0,u]}\{g(x)\} = g(0)$ (recall that g is continuous). Then thanks to Assumption B.2.(b) we can conclude that $\lim_{u\to 0}\frac{h(u)}{u^2} = \kappa g(0)$. Besides, this conclusion ensures that $\mu(\,\cdot\,,\{1\})$ is a finite measure.

- Finally, the mappings $u \mapsto \frac{h_n(u)}{1\wedge u^2}$ converge to $u \mapsto \frac{h(u)}{1\wedge u^2}$ uniformly on $\mathbb{R}^*$. Indeed, fix $\varepsilon > 0$. For all $u \in (0,1)$,

$$\begin{aligned}
\left|\frac{h_n(u)-h(u)}{u^2}\right| &\leq \frac{1}{u^2}\max_{[0,1]}|g| \int_0^u \Big(|f_n(nu)-f(u)|e^{-\tilde{\eta}_n y} + f(u)|e^{-\tilde{\eta}_n y} - e^{-\eta y}|\Big)\,\mathrm{d}y \\
&\leq \max_{[0,1]}|g| \left(\frac{|f_n(nu)-f(u)|}{u}\frac{1}{u}\int_0^u \mathrm{d}y + |\tilde{\eta}_n-\eta|\frac{1}{u^2}\int_0^u y\,\mathrm{d}y\right) \\
&\leq \max_{[0,1]}|g| \left(\frac{|f_n(nu)-f(u)|}{u} + \frac{1}{2}|\tilde{\eta}_n-\eta|\right).
\end{aligned}$$

  Then thanks to Assumption B.2.(a), and since $\tilde{\eta}_n \to \eta$, for n large enough $\left|\frac{h_n(u)-h(u)}{u^2}\right| \leq \varepsilon$ for all $u \in (0,1)$.
  On the other hand, for all $u \geq 1$,

$$|h_n(u)-h(u)| \leq \max_{[0,1]}|g|\,(|f_n(nu)-f(u)| + \frac{1}{2}|\tilde{\eta}_n-\eta|),$$

which can be upper bounded by ε for all $u \geq 1$ and n large enough, again thanks to Assumption B.2.(a) and to the convergence of $\tilde{\eta}_n$ towards η.
It follows then that for all $u \in \mathbb{R}^*$ and n large enough, $\left|\frac{h_n(u)}{1\wedge u^2}\right| \leq \varepsilon$.

All the conditions of Proposition 4.4 are then fulfilled, and we get the claimed convergence.

We now consider Assumption B.1. Exactly as before, we have

$$\begin{aligned}\Big|\int g(u)\mu_n(\mathrm{d}u,\{1\}) - \theta g(0)\Big| \leq \quad & \Big|\theta_n \int \tilde{\Lambda}_n(\mathrm{d}u)\Big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)\Big) - \theta g(0)\Big| \\ & + \Big|\theta_n \int \tilde{\Lambda}_n(\mathrm{d}u)\Big(\int_{1\wedge u}^{u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)\Big)\Big|,\end{aligned}$$

and as in case B.2, Lemma 4.10 entails the convergence to 0 of the second term in the right-hand side.

As for the first term, we have

$$\begin{aligned}&\Big|\theta_n \int \tilde{\Lambda}_n(\mathrm{d}u)\Big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(z)\Big) - \theta g(0)\Big| \\ &\qquad \leq \theta_n \int \tilde{\Lambda}_n(\mathrm{d}u)\Big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} |g(z)-g(0)|\Big) \\ &\qquad + \Big|\theta_n \int \tilde{\Lambda}_n(\mathrm{d}u)\Big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)} g(0)\Big) - \theta g(0)\Big|,\end{aligned}$$

Lemma 4.11 ensures the convergence to 0 of the second term in the right-hand side. Now the functions $u \mapsto \int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)}|g(z)-g(0)|$ are continuous, bounded by $\sup g$, and converge to $u \mapsto \int_0^{1\wedge u} \mathrm{d}z\, e^{\eta(z-u)}|g(z)-g(0)|$, which is equivalent to $g'(0)u^2/2$ as $u \to 0$. As a consequence of Proposition 4.4, we then have the convergence of $\int \tilde{\Lambda}_n(\mathrm{d}u)\big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)}|g(z)-g(0)|\big)$ to $g'(0)\frac{b^2}{2} + \int \tilde{\Lambda}(\mathrm{d}u)\big(\int_0^{1\wedge u} \mathrm{d}z\, e^{\eta(z-u)}|g(z)-g(0)|\big)$, which is a finite quantity, and thus the fact that $\theta_n \to 0$ ends the proof of the second assertion in (i).

The proof of point (ii) is very similar : Under Assumption B.1 or B.2, we have

$$\int g(u)\mu_n(\mathrm{d}u,\{0\}) = \int h_n(u)\ \tilde{\Lambda}_n(\mathrm{d}u)$$

with

$$h_n(u) := (1 - f_n(nu)) \int_0^u \mathrm{d}z\, e^{\tilde{\eta}_n(z-u)}\, g(z).$$

The same arguments as in the proof above work, except for the limit at 0 of $h(u)/u^2$: in this case, the fact that $g(u)/u \to g'(0)$ as $u \to 0$ implies $\frac{1}{u^2}\int_0^u e^{-\eta y} g(u-y)\mathrm{d}y \to \frac{g'(0)}{2}$, and then since $1 - f(u) \to 1$, we get $\frac{h(u)}{u^2} \to \frac{g'(0)}{2}$ as $u \to 0$. Finally, in the case of Assumption B.1, $f \equiv 0$ implies $\mu(\mathrm{d}u,\{0\}) = \mu^+(\mathrm{d}u)$, which allows us to conclude.

To get the last conclusion of the proposition, first notice that μ has no atom : Suppose μ has an atom $d > 0$, then $\mu(\{d\}) = \int_0^\infty e^{-\eta x}\Lambda(\{x+d\})\mathrm{d}x > 0$, which leads to the existence of a subset $U \subset [d,+\infty)$ such that $\mathrm{Leb}(U) \neq 0$ and $\Lambda(\{y\}) > 0$ for all $y \in U$. This implies $\Lambda(U) = +\infty$, which is impossible since $\Lambda(U) \leq \Lambda([d,+\infty)) < \infty$.

The results follow then by approximation : for all $\varepsilon > 0$, let I_ε^+ and I_ε^- be two continuous piecewise linear functions satisfying :

$$I_\varepsilon^+(x) = \begin{cases} 0 & \text{if} \quad x \leq \delta \\ 1 & \text{if} \quad x \geq \delta + \varepsilon \end{cases} \qquad I_\varepsilon^-(x) = \begin{cases} 0 & \text{if} \quad x \leq \delta - \varepsilon \\ 1 & \text{if} \quad x \geq \delta \end{cases} .$$

We have $I_\varepsilon^- \leq \mathbb{1}_{[0,\delta]} \leq I_\varepsilon^+$. This gives, for all $\varepsilon > 0$,

$$\int g I_\varepsilon^- \mathrm{d}\mu + \rho g(0) \leq \liminf_{n\to\infty} \int_{[0,\delta]} g \mathrm{d}\mu_n \leq \limsup_{n\to\infty} \int_{[0,\delta]} g \mathrm{d}\mu_n \leq \int g I_\varepsilon^+ \mathrm{d}\mu + \rho g(0).$$

Now when $\varepsilon \to 0$, $\int g I_\varepsilon^- \mathrm{d}\mu \to \int_{[0,\delta]} g \mathrm{d}\mu$ and $\int g I_\varepsilon^- \mathrm{d}\mu \to \int_{[0,\delta)} g \mathrm{d}\mu$. Since μ has no atom, these two integrals are equal and we get

$$\int_{[0,\delta]} g(u)\mu_n(\mathrm{d}u, \{1\}) \to \int_{[0,\delta]} g(u)\mu(\mathrm{d}u, \{1\}) + \rho g(0).$$

The other announced results can be obtained by a similar reasoning. □

Proof of Theorem 4.1 :
We first prove the second part of the theorem, i.e. we assume B.2. Moreover we assume first that Z does not drift to $-\infty$. Proposition 4.9 allows us to establish the three claims below, which correspond respectively to points (iii), (i) and (ii) of Proposition 0.8.

Claim 1: For all continuous bounded function g on $\mathbb{R}_+^2$ such that g is zero in a neighborhood of $(0,0)$,

$$\mu_n(g) \to (\mu + \rho\delta_{(0,1)})(g).$$

We have :

- First, since $u \mapsto g(u,0)$ is zero in a neighborhood of 0,

$$\int g(u,0)\mu_n(\mathrm{d}u, \{0\}) \to \int g(u,0)\mu(\mathrm{d}u, \{0\})$$

  as $n \to \infty$ thanks to Proposition 4.9 (ii).

- Second $\int g(u,1)\mu_n(\mathrm{d}u, \{1\}) \to \int g(u,1)\mu(\mathrm{d}u, \{1\}) + \rho g(0,1)$ according to Proposition 4.9 (i).

and the result follows.

Claim 2: For all $(\alpha, \beta) \in \mathbb{R}_+^2$,

$$\int (\alpha,\beta)^t h(u,q)\mu_n(\mathrm{d}u, \mathrm{d}q) \to \frac{b^2}{2}\alpha + \rho\beta + \int (\alpha,\beta)^t h(u,q)\mu(\mathrm{d}u, \mathrm{d}q),$$

where h is the truncation function defined earlier.

We have

$$\int (\alpha,\beta)^t h(u,q)\mu_n(\mathrm{d}u, \mathrm{d}q) = \int_{[0,\delta]} (\alpha u + \beta q)\mu_n(\mathrm{d}u, \mathrm{d}q) + \int_{(\delta,\infty)} (\alpha\delta + \beta q)\mu_n(\mathrm{d}u, \mathrm{d}q),$$

and then :

- $u \mapsto \alpha u + \beta$ is a continuous bounded function on $[0, \delta]$, then thanks to Proposition 4.9,

$$\int_{[0,\delta]} (\alpha u + \beta)\mu_n(\mathrm{d}u, \{1\}) \to \rho\beta + \int_{[0,\delta]} (\alpha u + \beta)\mu(\mathrm{d}u, \{1\}).$$

- In the same way, thanks to Proposition 4.9 (ii),

$$\int_{[0,\delta]} \alpha u \mu_n(\mathrm{d}u, \{0\}) \to \frac{b^2}{2}\alpha + \int_{[0,\delta]} \alpha u \mu(\mathrm{d}u, \{0\}).$$

- And finally, thanks to Proposition 4.9 (points (i) and (ii)),

$$\int_{(\delta,\infty)} (\alpha\delta + \beta q)\mu_n(\mathrm{d}u, \mathrm{d}q) \to \int_{(\delta,\infty)} (\alpha\delta + \beta q)\mu(\mathrm{d}u, \mathrm{d}q) \quad \text{when } n \to \infty.$$

As a consequence,

$$\begin{aligned}
&\int (\alpha, \beta)^t h(u, q)\mu_n(\mathrm{d}u, \mathrm{d}q) \\
&\to \rho\beta + \int_{[0,\delta]} (\alpha u + \beta)\mu(\mathrm{d}u, \{1\}) + \frac{b^2}{2}\alpha + \int_{[0,\delta]} \alpha u \mu(\mathrm{d}u, \{0\}) + \int_{(\delta,\infty)} (\alpha\delta + \beta q)\mu(\mathrm{d}u, \mathrm{d}q) \\
&= \rho\beta + \frac{b^2}{2}\alpha + \int (\alpha, \beta)^t h(u, q)\mu(\mathrm{d}u, \mathrm{d}q),
\end{aligned}$$

which proves our assertion.

Claim 3: Denote by h_1 (resp. h_2) the first (resp. second) coordinate of h. For all $i, j \in \{1, 2\}$,

$$\int h_i(u, q)h_j(u, q)\mu_n(\mathrm{d}u, \mathrm{d}q) \underset{n\to\infty}{\to} \int h_i(u, q)h_j(u, q)(\mu(\mathrm{d}u, \mathrm{d}q) + \rho\delta_0(\mathrm{d}u)\delta_1(\mathrm{d}q))$$

as $n \to \infty$.

Note that $\int h_i(u, q)h_j(u, q)\delta_0(\mathrm{d}u)\delta_1(\mathrm{d}q) = h_i(0, 1)h_j(0, 1)$.

- The continuous bounded function h_1^2 satisfies $h_1(u, q)^2/u \to 0$ as $u \to 0$, for $q \in \{0, 1\}$. Then thanks to Proposition 4.9 (points (i) and (ii)) we have

$$\int h_1(u, q)^2\mu_n(\mathrm{d}u, \mathrm{d}q) \underset{n\to\infty}{\to} \int h_1(u, q)^2\mu(\mathrm{d}u, \mathrm{d}q),$$

  and since $h_1(0, 1) = 0$, we get the announced result for $(i, j) = (1, 1)$.

- The continuous bounded function $u \mapsto h_1(u, 1)h_2(u, 1)$ satisfies $h_1(0, 1)h_2(0, 1) = 0$ as $u \to 0$, so that according to Proposition 4.9 (i),

$$\int h_1(u, 1)h_2(u, 1)\mu_n(\mathrm{d}u, \{1\}) \underset{n\to\infty}{\to} \int h_1(u, 1)h_2(u, 1)\mu(\mathrm{d}u, \{1\}).$$

  Moreover, $h_1(0, 1) = 0$ and $h_2(u, 0) = 0$ for all $u \geq 0$, and then we can deduce the result for $(i, j) = (1, 2)$.

- Finally, when $q = 0$ or $q = 1$, we have $h_2(u, q)^2 \equiv q$ for all $u \in \mathbb{R}_+$. In consequence,

$$\int h_2(u,1)^2 \mu_n(\mathrm{d}u, \{1\}) \underset{n\to\infty}{\to} \int h_2(u,1)^2 \mu(\mathrm{d}u, \{1\}) + \rho h_2(0,1)^2,$$

and since $h_2(u, 0) \equiv 0$, we get the result for $(i, j) = (2, 2)$.

Finally the three claims establish the theorem under Assumption B.2 through a straightforward application of Proposition 0.8. The proof in the case of Assumption B.1 is very similar, and since in this case $f \equiv 0$, the limiting Lévy measure is

$$\mu(\mathrm{d}u, \{0, 1\}) + \theta\delta(\mathrm{d}q)\delta_0(\mathrm{d}u) = \mu^+(\mathrm{d}u)\delta_0(\mathrm{d}q) + \theta\delta_1(\mathrm{d}q)\delta_0(\mathrm{d}u),$$

which gives the expected result.

Finally we prove the theorem in the case where Z drifts to $-\infty$. Using the convention that an exponentially distributed variable with parameter 0 is equal to $+\infty$ a.s., and setting $k_n := 0$ when $\tilde{Z}_n$ does not drift to $-\infty$, all that is needed now is to prove that $k_n \to k$ as $n \to \infty$. Now since $W(\infty) < +\infty$, from the uniform convergence on $\mathbb{R}_+$ of $\tilde{W}_n$ towards W (Proposition 4.3), we have $\tilde{W}_n(\infty) \to W(\infty)$, which ends the proof. □

5 Joint convergence in distribution of $\tilde{Z}_n$ with its local time at the supremum and its marked ladder height process

In this section we assume that Assumption A, and one of the two Assumptions B.1 or B.2 hold, and we establish the joint convergence in law of $(\tilde{Z}_n, L_n, H_n^+, H_n^{\mathrm{M}})$. To prove this result, we will need the convergence in distribution of H_n^- established in Section 4, and the joint convergence in distribution of $\tilde{Z}_n$ with its local time at the supremum and its classical ladder height process. The latter convergence is proved in L. Chaumont and R.A. Doney [CD10], in the case of Lévy processes for which 0 is regular for the open half-line $(0, \infty)$. We adapt here their proofs to our case of spectrally positive Lévy processes with finite variation.

Theorem 5.1. *The following convergence in distribution holds in $\mathbb{D}(\mathbb{R})^4$ as $n \to \infty$:*

$$(\tilde{Z}_n, L_n, H_n^+, H_n^-, H_n^{\mathrm{M}}) \Rightarrow (Z, L, H^+, H^-, H^{\mathrm{M}}),$$

where conditional on (Z, L, H^+, H^-), H^{M} is a Poisson process whose jump process is the jump process of $H^+ + H^-$.

This theorem is a consequence of the following proposition :

Proposition 5.2. *We have the following joint convergence in distribution in $\mathbb{D}(\mathbb{R})^4$ as $n \to \infty$:*

$$(\tilde{Z}_n, L_n, H_n^+, H_n^-) \Rightarrow (Z, L, H^+, H^-).$$

Proof of Theorem 5.1 :
Consider the process $(H_n^+ + H_n^-)$, denote by $\pi^{\pm}$ its jump point process (with values in $\mathbb{R}_+^* \times \{\partial\}$), and define $A := \{t \in \mathbb{R}_+,\ \pi^{\pm}(t) \in \mathbb{R}_+^*\}$. Then we define the random process π^{M} as follows : conditional on $(H_n^+ + H_n^-)$, for any t in the countable set A, $\pi^{\mathrm{M}}(t)$ follows a Bernoulli distribution with parameter $\mathbb{B}_{f_n(\pi^{\pm}(t))}$, and for $t \notin A$, $\pi^{\mathrm{M}}(t) = \partial$. Then by definition the process H_n^{M} is distributed as a Poisson process with jump point process π^{M}. It follows that conditional on (H_n^+, H_n^-), the process H_n^{M} is independent of $\tilde{Z}_n$ and L_n. Then Theorem 4.1 along with Proposition 5.2 entail

the joint convergence in distribution of $(\tilde{Z}_n, L_n, H_n^+, H_n^-, H_n^{\text{M}})$ towards $(Z, L, H^+, H^-, H^{\text{M}})$, and Theorem 5.1 follows. □

We now want to prove Proposition 5.2, for which our inspiration comes from L. Chaumont and R.A. Doney [CD10]. With this aim in view, we need to introduce some notions about random walks. We consider the random walk $S = (S(j))_{j\geq 0}$ defined by $S(0) = 0$ and $S(j) = \sum_{i=1}^{j} Y_i$ for $j \geq 1$, where $(Y_i)_{i\geq 1}$ is a sequence of i.i.d. $\mathbb{R}$-valued random variables. We endow our random walk S with a sequence of i.i.d. exponential random variables $(a_i)_{i\geq 1}$ (their common parameter can be chosen arbitrarily), independent of S. We write $(N_t)_{t\geq 0}$ for the Poisson process associated with this sequence of variables. We denote by $\bar{S}(j)$ the maximum of the random walk at step j : $\bar{S}(j) := \max\{S(i),\ 1 \leq i \leq j\}$, and we define its local time at the maximum :

$$\mathtt{k}(j) := \#\{i \in \{1, \dots, j\},\ S(i) > \bar{S}(i-1)\}.$$

We then introduce a continuous-state version of the local time of S at its maximum by setting

$$K(j) := \sum_{i=1}^{\mathtt{k}(j)} a_i.$$

We denote by $\mathtt{t}$ the right inverse of $\mathtt{k}$:

$$\mathtt{t}(0) = 0,\ \mathtt{t}(j+1) = \min\{i > \mathtt{t}(j),\ S(i) > S(\mathtt{t}(j))\},$$

which implies $\mathtt{k}(\mathtt{t}(j)) = j$ for all integer j. Then similarly for K, we define T by

$$\forall s \geq 0,\ T(s) = \inf\{h \geq 0, K(h) > s\},$$

which satisfies $T = \mathtt{t} \circ N$. Finally, we define $\mathtt{g}$ and G as follows :

$$\forall j \geq 0,\ \mathtt{g}(j) = \bar{S}(\mathtt{t}(j)), \quad \text{and } \forall s \geq 0,\ G(s) = \bar{S}(T(s)).$$

The pair of processes $(\mathtt{t}, \mathtt{g})$ is called ladder process, $\mathtt{t}$ being the ladder time process, and $\mathtt{g}$ the ladder height process. The pair (T, G) is then a continuous-time version of the classical ladder process $(\mathtt{t}, \mathtt{g})$.

In the sequel, we will consider a sequence of random walks $(S_n)_{n\geq 1}$ (whose distributions can depend on n). As before, and independently for all n, we endow the random walk S_n with a sequence of i.i.d. exponential variables $(A_i^n)_{i\geq 1}$, independent of S_n, with parameter α_n to be specified later, and we denote by N^n the corresponding Poisson process. We will use an obvious notation with subscript n for all the quantities involved by the random walk S_n.

Let X be a spectrally positive Lévy process (which is not a subordinator) with finite variation. We define its local time L_X as in Section 0.1.3 :

$$L_X(t) := \sum_{i=0}^{\mathtt{l}(t)} A_i,$$

where $\mathtt{l}(t)$ represents the number of jumps of the supremum until time t, and $(A_i)_{i\geq 0}$ is a sequence of i.i.d. random exponential variables with arbitrarily chosen parameter α, independent from X. We denote by (L_X^{-1}, H) its bivariate ladder process and by κ the Laplace exponent of the latter.

We define the convergence in distribution (resp. a.s.) of the sequence (S_n) towards X to be equivalent to the convergence in distribution (resp. a.s.) of the sequence of continuous-time processes $(S_n[nt])_{t\geq 0}$ towards X, in $\mathbb{D}(\mathbb{R}_+)$. We keep again the notation $S_n \Rightarrow X$ for the convergence in law of S_n to X.

The following four statements are the respective analogues of Theorem 1, Theorem 2, Theorem 3 and Corollary 2 in [CD10], in the case of Lévy processes for which 0 is not regular for the open half-line $(0,\infty)$. Our proofs are widely inspired of that of Chaumont and Doney in this paper.

Proposition 5.3. *Let (S_n) be a sequence of random walks converging in distribution to the Lévy process X. We then have the following convergence in law :*

$$\left(\frac{1}{n}T_n, G_n\right) \Rightarrow (L_X^{-1}, H),$$

where for all n, the parameter α_n of the Poisson process N^n is given by

$$\alpha_n := \exp\{\sum_{k\geq 1}\frac{1}{k}e^{-k/n}\mathbb{P}(S_n(k)>0)\}.$$

Proof :
The key of the following calculation is Fristedt's formula, which can be found in [Don07, th. 10] :

$$1-\mathbb{E}(e^{-\delta \mathtt{t}_n(1)-\beta \mathtt{g}_n(1)}) = \exp\{-\sum_{k\geq 1}\frac{e^{-\delta k}}{k}\mathbb{E}(e^{-\beta S_n(k)},\ S_n(k)>0)\}.$$

It allows us to calculate the Laplace transform of $(\frac{1}{n}T_n, G_n)$ for all $\delta,\beta>0$:

$$\begin{aligned}
\mathbb{E}(e^{-\delta T_n(1)-\beta G_n(1)}) &= \mathbb{E}(e^{-\delta \mathtt{t}_n(N_1^n)-\beta \mathtt{g}_n(N_1^n)}) \\
&= \sum_{j\geq 0}\mathbb{E}(e^{-\delta \mathtt{t}_n(1)-\beta \mathtt{g}_n(1)})^j\mathbb{P}(N_1^n=j) \\
&= e^{-\alpha_n}\sum_{j\geq 0}\left(1-\exp\left\{-\int_{1/n}^{\infty}\frac{n}{[nt]}e^{-\delta[nt]/n}\mathbb{E}(e^{-\beta S_n([nt])},\ S_n([nt])>0)\mathrm{d}t\right\}\right)^j\frac{(\alpha_n)^j}{j!} \\
&= \exp\left\{-\alpha_n\exp\left(-\int_{1/n}^{\infty}\frac{n}{[nt]}e^{-\delta[nt]/n}\mathbb{E}(e^{-\beta S_n([nt])},\ S_n([nt])>0)\mathrm{d}t\right)\right\}.
\end{aligned}$$

Now from the expression of α_n we have

$$\alpha_n = \exp\left(\int_{1/n}^{\infty}\frac{n}{[nt]}e^{-[nt]/n}\mathbb{P}(S_n([nt])>0)\mathrm{d}t\right), \tag{7}$$

and the convergence of S_n towards X gives, with an argument of dominated convergence as in the proof of Proposition 4.7,

$$\begin{aligned}
&\alpha_n\exp\left(-\int_{1/n}^{\infty}\frac{n}{[nt]}e^{-\delta[nt]/n}\mathbb{E}(e^{-\beta S_n([nt])},\ S_n([nt])>0)\mathrm{d}t\right) \\
&\qquad\underset{n\to\infty}{\to}\exp\left(-\int_0^{\infty}\left(\frac{e^{-t}}{t}\mathbb{P}(X_t>0)-\frac{e^{-\delta t}}{t}\mathbb{E}(e^{-\beta X_t},\ X_t>0)\right)\mathrm{d}t\right) \\
&\qquad\quad=\exp\left(-\int_0^{\infty}\frac{1}{t}\mathbb{E}(e^{-t}-e^{-\delta t-\beta X_t},\ X_t>0)\mathrm{d}t\right) \\
&\qquad\quad=\kappa(\delta,\beta),
\end{aligned}$$

according to Corollary VI.10 in [Ber96].
Thus we get the convergence of the Laplace exponent of $(\frac{1}{n}T_n, G_n)$ towards that of (L_X^{-1}, H), which ends the proof. □

Corollary 5.4. *The parameters α_n converge to α as $n \to \infty$.*

Proof :
We saw in the proof above (see formula (7) and following computation) that as $n \to \infty$,

$$\alpha_n \to \exp\Big\{ \int_0^\infty \frac{e^{-t}}{t} \mathbb{P}(X_t > 0) \mathrm{d}t \Big\}.$$

Now this quantity is equal to $\kappa(\infty, 0) := \lim_{\delta \to \infty} \kappa(\delta, 0)$, and we have

$$\exp(\kappa(\infty, 0)) = \lim_{\delta \to \infty} \mathbb{E}(e^{-\delta L_X^{-1}(1)}) = \mathbb{P}(A_1 > 1) = e^{-\alpha}.$$

□

Proposition 5.5. *Under the same statement as in Proposition 5.3, assuming furthermore that the convergence of (S_n) towards X holds almost surely, for all fixed $t \geq 0$ we have the convergence in probability of $K_n([nt])$ towards $L_X(t)$.*

Proof :
Fix $\varepsilon > 0$ and $t \geq 0$. Recall from the definitions of K_n and L_X that for all $n \geq 1$, $j \geq 0$, $t \geq 0$,

$$K_n(j) = \sum_{i=1}^{\mathtt{k}_n(j)} A_i^n \quad \text{and} \quad L_X(t) = \sum_{i=1}^{\mathtt{l}(t)} A_i.$$

Write

$$\mathbb{P}\left(\Big| \sum_{i=1}^{\mathtt{k}_n([nt])} A_i^n - \sum_{i=1}^{\mathtt{l}(t)} A_i \Big| > \varepsilon \right) \leq \mathbb{P}\left(|\mathtt{k}_n([nt]) - \mathtt{l}(t)| > 0 \right) + \mathbb{P}\left(\sum_{i=1}^{\mathtt{l}(t)} |A_i - A_i^n| > \varepsilon \right).$$

Fix $\eta > 0$. On the one hand, since $\mathtt{k}_n$ and $\mathtt{l}$ are finite integers, the almost sure convergence of (S_n) towards X ensures that for all $t \geq 0$, $\mathtt{k}_n([nt]) \to \mathtt{l}(t)$ a.s. , and consequently for n large enough

$$\mathbb{P}(|\mathtt{k}_n([nt]) - \mathtt{l}(t)| > 0) \leq \frac{\eta}{3}.$$

On the other hand, thanks to Corollary 5.4 we can find $u > 0$ and $n_0 \geq 0$ such that for $n \geq n_0$, $\mathbb{P}(\mathtt{l}^{-1}(u) < t) < \eta/3$, and $\frac{u}{\varepsilon}\Big|\frac{1}{\alpha_n} - \frac{1}{\alpha}\Big| < \frac{\eta}{3}$, where $\mathtt{l}^{-1}(u) := \inf\{s \geq 0,\ \mathtt{l}(s) > u\}$ denotes the right inverse of $\mathtt{l}$. Then for $n \geq n_0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{\mathtt{l}(t)} |A_i - A_i^n| > \varepsilon \right) &\leq \mathbb{P}\left(\sum_{i=1}^{u} |A_i - A_i^n| > \varepsilon \right) + \mathbb{P}(\mathtt{l}^{-1}(u) < t) \\
&\leq \frac{\mathbb{E}(\sum_1^u |A_i^n - A_i|)}{\varepsilon} + \frac{\eta}{3} \\
&\leq \frac{u}{\varepsilon}\Big|\frac{1}{\alpha_n} - \frac{1}{\alpha}\Big| + \frac{\eta}{3} \\
&\leq \frac{2\eta}{3},
\end{aligned}$$

where the second inequality is obtained from an appeal to Markov's inequality. We conclude that $\lim_{n\to\infty} \mathbb{P}(|K_n[nt] - L_X(t)| > \varepsilon) = 0$. □

Next let us turn our attention back to our sequence of spectrally positive Lévy processes $(\tilde{Z}_n)$ converging to a Lévy process Z with infinite variation.

Proposition 5.6. *If the convergence of $\tilde{Z}_n$ to Z holds a.s., then for all $t \geq 0$, we have convergence in probability of $L_n(t)$ towards $L(t)$.*

Proof :
As in [CD10], for all $n \geq 0$, we consider the sequence of random walks $(S_{n,k})_{k\geq 0}$ defined by $S_{n,k}(j) = \tilde{Z}_n(j/k)$ for all $j \geq 0$, so that as $k \to \infty$,

$$(S_{n,k}([kt]))_{t\geq 0} \to \tilde{Z}_n \text{ a.s.}$$

As previously, each random walk $S_{n,k}$ is endowed, independently of the others, with a Poisson process $N^{n,k}$ with parameter $\alpha_{n,k} := \exp\{\sum_{i\geq 1} \frac{1}{i} e^{-i/k} \mathbb{P}(S_{n,k}(i) > 0)\}$. We will use the obvious notation with subscript n, k for all the quantities defined earlier involved by $S_{n,k}$.
Fix $\varepsilon > 0$. From Proposition 5.5, we can find some sequence of integers $(k_n)_{n\geq 1}$ such that, as $n \to \infty$,

$$(S_{n,k_n}([k_n t]))_{t\geq 0} \to Z \text{ a.s.}$$

and

$$\mathbb{P}(|K_{n,k_n}[k_n t] - L_n(t)| > \varepsilon) \to 0.$$

We have

$$\begin{aligned}
\mathbb{P}(|L_n(t) - L(t)| > 3\varepsilon) \leq\ & \mathbb{P}(|L_n(t) - K_{n,k_n}[k_n t]| > \varepsilon) \\
& + \mathbb{P}(|K_{n,k_n}[k_n t] - \frac{1}{\alpha_{n,k_n}} \mathtt{k}_{n,k_n}([k_n t])| > \varepsilon) \\
& + \mathbb{P}(|\frac{1}{\alpha_{n,k_n}} \mathtt{k}_{n,k_n}([k_n t]) - L_t| > \varepsilon).
\end{aligned}$$

We chose the subsequence (k_n) such that the first term in the sum goes to 0 as $n \to \infty$. The a.s. convergence of (S_{n,k_n}) towards the Lévy process Z, for which the state 0 is regular for $(0, \infty)$, allows us to apply Theorem 2 in [CD10] to get the convergence towards 0 of the last term in the sum.
It remains to prove that $K_{n,k_n}[k_n t] - \frac{1}{\alpha_{n,k_n}} \mathtt{k}_{n,k_n}([k_n t])$ converges in probability to 0 as $n \to \infty$. Recall that for all $n, j \geq 0$, $\mathtt{k}_{n,k_n}(\mathtt{t}_{n,k_n}(j)) = j$. Thus for all $j \geq 0$, we can write

$$\begin{aligned}
&\mathbb{P}(|K_{n,k_n}[k_n t] - \frac{1}{\alpha_{n,k_n}} \mathtt{k}_{n,k_n}([k_n t])| > \varepsilon) \\
&\quad = \mathbb{P}\left(\sum_{i=1}^{\mathtt{k}_{n,k_n}[k_n t]} \left| A_i^{n,k_n} - \frac{1}{\alpha_{n,k_n}} \right| > \varepsilon \right) \\
&\quad \leq \mathbb{P}\left(\sum_{i=1}^{[\alpha_{n,k_n} j t]} \left| A_i^{n,k_n} - \frac{1}{\alpha_{n,k_n}} \right| > \varepsilon \right) + \mathbb{P}(\mathtt{t}_{n,k_n}[\alpha_{n,k_n} j t] < k_n t) \\
&\quad \leq \frac{[\alpha_{n,k_n} j t]}{\varepsilon^2 \alpha_{n,k_n}^2} + \mathbb{P}(\mathtt{t}_{n,k_n}[\alpha_{n,k_n} j t] < k_n t),
\end{aligned}$$

the last inequality coming from the Bienaymé-Tchebitchev's inequality. From Remark 1 in [CD10], we know that $\lim_{n\to\infty} \alpha_{n,k_n} = +\infty$. Thus letting first n tend to ∞, we have that $\frac{[\alpha_{n,k_n} jt]}{\varepsilon^2 \alpha^2_{n,k_n}}$ goes to 0, and $\mathbb{P}(\mathtt{t}_{n,k_n}[\alpha_{n,k_n} jt] < k_n t)$ tends to $\mathbb{P}(L^{-1}(jt) < t)$ according to Theorem 1 in [CD10]. This last quantity now goes to 0 as $j \to \infty$, and we completed the proof. □

Corollary 5.7. *The sequence $(\tilde{Z}_n, L_n, L_n^{-1}, H_n^+)$ converges as $n \to \infty$, in the sense of the finite dimensional distributions, to the process (Z, L, L^{-1}, H^+).*

Proof :
By Skorokhod's representation we may suppose that the convergence of $\tilde{Z}_n$ towards Z holds a.s. Now Proposition 4.7 and Theorem 4.1 ensure the convergence in law of each coordinate, which provides the tightness of the quadruplet. Then, proving the a.s. convergence of the finite dimensional marginals will be sufficient to establish the corollary.
Now fix $t > 0$. From Proposition 5.6 we know that there exists some sequence of integers k_n, going to ∞ as $n \to \infty$, such that $L_{k_n}(t)$ tends to $L(t)$ a.s. From the definition of the inverse local time as a first passage time, and noting that L has no fixed time of discontinuity, the latter convergence implies that of $L_{k_n}^{-1}(t)$ to $L^{-1}(t)$ a.s., by virtue of Proposition VI.2.11 in [JS87].

As said in [CD10], $L^{-1}(t)$ is an announceable stopping time (here 0 is regular for Z for $(0, \infty)$), so that from an appeal to Exercise 3 in [Ber96], we get that Z is a.s. continuous at time $L^{-1}(t)$. According to VI.2.3 in [JS87], for all (possibly random) continuity point u of Z, we have $\tilde{Z}_{k_n}(u) \to Z(u)$ a.s. as $n \to \infty$, and hence the sequence $(\tilde{Z}_{k_n}(t), L_{k_n}(t), L_{k_n}^{-1}(t), \tilde{Z}_{k_n}(L_{k_n}^{-1}(t)))$ converges a.s. as $n \to \infty$ towards $(Z(t), L(t), L^{-1}(t), H^+(t))$.

Finally, taking any sequence of times $t_1 < t_2 < \ldots < t_j$, we can find a sequence k'_n of integers tending to ∞ as $n \to \infty$, such that $\big((\tilde{Z}_{k'_n}(t_i), L_{k'_n}(t_i), L_{k'_n}^{-1}(t_i)), H_{k'_n}^+(t_i), 1 \le i \le j\big)$ converges a.s. towards $\big((Z(t_i), L(t_i), L^{-1}(t_i), H^+(t_i)), 1 \le i \le j\big)$ as $n \to \infty$, which ends the proof. □

Proof of Proposition 5.2 :
According to Assumption A (resp. Proposition 5.6, Proposition 4.7, Remark 4.2), we know that $\tilde{Z}_n$ (resp. L_n, L_n^{-1}, H_n^+, H_n^-) converges in distribution towards Z (resp. L, L^{-1}, H^+, H^-). Therefore, each of these sequences is tight, and in consequence the sequence $(\tilde{Z}_n, L_n, L_n^{-1}, H_n^+, H_n^-)$ is tight. From Corollary 5.7 we then get the joint convergence in distribution of $(\tilde{Z}_n, L_n, L_n^{-1}, H_n^+)$ towards (Z, L, L^{-1}, H^+), and moreover, the tightness ensures the existence of a subsequence (k_n) such that $(\tilde{Z}_{k_n}, L_{k_n}, L_{k_n}^{-1}, H_{k_n}^+, H_{k_n}^-)$ converges in distribution to $(\tilde{Z}, \tilde{L}, \tilde{L^{-1}}, \tilde{H}^+, \tilde{H}^-)$, with $(\tilde{Z}, \tilde{L}, \tilde{L}^{-1}, \tilde{H}^+) \overset{\mathcal{L}}{=} (Z, L, L^{-1}, H^+)$ and $\tilde{H}^- \overset{\mathcal{L}}{=} H^-$. By virtue of the Skorokhod representation theorem, we can suppose that this convergence holds a.s.

The processes $\tilde{H}^-$ and H^- are two subordinators and are equal in law, thus their continuous parts (which are deterministic drifts) are equal in law and therefore, almost surely. Consider now the jump part of $\tilde{H}^-$. For all $\varepsilon > 0$ and $y \in \mathbb{D}(\mathbb{R}_+)$, define

$$U(y, \varepsilon) := \{u > 0,\ |\Delta y(t)| = u \text{ for some } t\},$$

and

$$t^0(y, \varepsilon) := 0, \text{ and } \forall p \ge 0,\ t^{p+1}(y, \varepsilon) := \inf\{t > t^p(y, \varepsilon),\ |\Delta y(t)| > \varepsilon\}.$$

Proposition VI.2.7 in [JS87] ensures for all $p \ge 0$ that the mapping $y \mapsto t^p(y, \varepsilon)$ (resp. $y \mapsto \Delta y(t^p(y, \varepsilon))$) is continuous on $\mathbb{D}(\mathbb{R}_+)$ at each point y such that $\varepsilon \notin U(y, \varepsilon)$ (resp. $\varepsilon \notin U(y, \varepsilon)$

and $t^p(y,\varepsilon) < \infty$).
Now we know that $t^p(Z,\varepsilon)$, $t^p(H^+,\varepsilon)$, $t^p(H^-,\varepsilon)$ are finite a.s., and μ^+, μ^- have no atoms. Moreover, since Λ is a σ-finite measure on $\mathbb{R}_+^*$, there exists a sequence $(\varepsilon_m)_{m\geq 1}$ of positive real numbers, which vanishes as $m \to \infty$, such that $\Lambda(\{\varepsilon_m\}) = 0$. As a consequence, for all $m \geq 1$, the functions $y \mapsto t^p(y,\varepsilon_m)$ and $y \mapsto \Delta y(t^p(y,\varepsilon_m))$ are a.s. continuous w.r.t the distribution of Z, H^+ and H^-. Along with Proposition VI.2.1 of [JS87], this gives for all $m \geq 1$ the following almost sure convergence as $n \to \infty$:

$$\begin{aligned}(t^p(H^+_{k_n},\varepsilon_m), \Delta H^+_{k_n}(t^p(H^+_{k_n},\varepsilon_m)), \Delta\tilde{Z}_{k_n}(L^{-1}_{k_n}(t^p(H^+_{k_n},\varepsilon_m))), \Delta H^-_{k_n}(t^p(H^+_{k_n},\varepsilon_m)))\\ \xrightarrow{a.s.} (t^p(\tilde{H}^+,\varepsilon_m), \Delta\tilde{H}^+(t^p(\tilde{H}^+,\varepsilon_m)), \Delta\tilde{Z}(\tilde{L}^{-1}(t^p(\tilde{H}^+,\varepsilon_m))), \Delta\tilde{H}^-(t^p(\tilde{H}^+,\varepsilon_m))).\end{aligned}$$

Now with probability one the jumping times of H_n^+ are exactly those of H_n^-, and for all $t > 0$, $\Delta H_n^-(t) = \Delta\tilde{Z}_n(L_n^{-1}(t)) - \Delta H_n^+(t)$ a.s. Therefore letting now $m \to \infty$, we get :

$$\sum_{s\leq t} \Delta\tilde{H}^-(s) = \sum_{s\leq t} (\Delta\tilde{Z}(\tilde{L}^{-1}(s)) - \Delta\tilde{H}^+(s)) \quad \text{a.s.}$$

As a consequence, we have $(\tilde{Z},\tilde{L},\tilde{L}^{-1},\tilde{H}^+,\tilde{H}^-) \overset{\mathcal{L}}{=} (Z,L,L^{-1},H^+,H^-)$, and then we get the convergence in distribution of $(\tilde{Z}_n, L_n, L_n^{-1}, H_n^+, H_n^-)$ towards (Z,L,L^{-1},H^+,H^-). □

Chapter II

Lévy processes with marked jumps II : Invariance principle for branching processes with mutations

The article [Del13b] is submitted to Stochastic Processes and their Applications.

1 Introduction

A splitting tree [Gei96, GK97, Lam10] describes a population of individuals with i.i.d. lifetime durations, which distribution is not necessarily exponential, giving birth at constant rate during their lives. Each birth gives rise to a single child, who behaves as an independent copy of her parent. We consider here the extended framework of [Lam10] : for each individual, the birth times and lifetimes of her progeny is given by a Poisson process with intensity $\mathrm{d}t \cdot \Lambda(\mathrm{d}r)$, where the so-called lifespan measure Λ is a Lévy measure on $(0, \infty)$ satisfying $\int (1 \wedge r)\Lambda(\mathrm{d}r) < \infty$. In particular, the number of children of a given individual is possibly infinite. In addition, we assume that individuals carry types, and that every time a birth occurs, a mutation may happen, giving rise to a mutant child. Mutations are assumed to be neutral, meaning that they do not affect the behaviour of individuals. In order to take this into account, we introduce *marked splitting trees* : to each birth event we associate a mark in $\{0, 1\}$, which will code for the absence (0) or presence (1) of a mutation. In other words, a 0-type birth means a clonal birth, and a 1-type birth produces a mutant child. The mutations experienced by the population are then described by these marks.

Population models with mutations have inspired lots of works in the past, and have many applications in domains such as population genetics, phylogeny or epidemiology. Such models have been well studied in the particular case of populations with fixed size. In the Wright-Fisher and Moran models with neutral mutations, as well as in the Kingman coalescent, explicit results on the allelic partition of the population are provided by Ewens' sampling formula [Ewe72, Dur08]. Relaxing the hypotheses of constant population size, branching processes with mutations at birth are studied in the monography [Taï92]. More recently, results have been obtained for the allelic partition and frequency spectrum of splitting trees, with mutations appearing either at birth of individuals [Ric14] or at constant rate along the lineages [Lam11, CL12a, CL12b],

and are reviewed in [CLR12]. The present work focuses on asymptotic results when the size of the population gets large, for the genealogy (with mutational history) of splitting trees with mutations at birth, and relies on a previous article [Del13a] (Chapter I in this thesis).

Genealogy of the n-th population Let us fix some positive real number τ. For $n \in \mathbb{N}$, consider a marked splitting tree $\mathbb{T}_n$, and condition it on having a fixed positive number I_n of individuals alive at level τ. Note that we use here the word 'level' to denote the real time in which the individuals live, whereas we reserve the word 'time' for the index of stochastic processes. This paper follows on from a work of L. Popovic [Pop04] in the critical case with exponential lifetimes, without mutations, in which she proved the convergence in distribution of the coalescent point process (i.e. the smallest subtree containing the genealogy of the extant individuals) towards a certain Poisson point process. Our aim is to provide asymptotic results as I_n gets large, for the structure of the genealogy of the population up to level τ, enriched with the random levels at which marks occurred on the lineages. To this aim, after a proper rescaling of $\mathbb{T}_n$, we introduce a random point measure Σ_n which we call the marked coalescent point process. This point measure has $I_n - 1$ atoms; its i-th one is itself a random point measure, whose set of atoms contains all the levels where mutations occurred on the i-th lineage, and the coalescence time between individuals i and $i-1$. This sequence of point measures (Σ_n) is the mathematical object for which we aim to get convergence as $n \to \infty$, after having set some convergence assumptions, which we discuss later.

Our work mainly relies on the study of splitting trees with the help of the so-called jumping chronological contour process (or JCCP). This process is an exploration process of the tree (without mutations) introduced by A. Lambert in [Lam10], visiting all the existence levels of all the individuals exactly once, and ending at level 0. He showed in this paper that the JCCP of a tree truncated up to level τ is a compensated compound Poisson process with no negative jumps (spectrally positive Lévy process with finite variation) reflected below τ and killed when hitting 0. In particular, the labeling of the excursions of the JCCP below τ provides a labeling of the extant individuals at level τ. Inferring properties concerning the genealogy of the alive population at level τ in the tree then essentially consists in studying the excursions away from τ of this reflected Lévy process.

We introduce in Chapter I a generalization of this contour process to the framework of our rescaled marked splitting trees $(\tilde{\mathbb{T}}_n)$. We are thereby led to study a bivariate Lévy process $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$. Roughly speaking, $\tilde{Z}_n$ codes for the JCCP of $\tilde{\mathbb{T}}_n$ (without mutations), and $\tilde{Z}_n^{\mathrm{M}}$ codes for the mutations. Namely, since a jump of $\tilde{Z}_n$ corresponds to the encounter of a birth event when exploring the tree, $\tilde{Z}_n^{\mathrm{M}}$ will jump as well (with amplitude 1) if this birth was of type 1. The process $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ is in one-to-one correspondence with the marked tree $\tilde{\mathbb{T}}_n$. We now want to characterize the law of the atoms of Σ_n using this property. Let us first give an idea of our reasoning in the case where there is no mutations. The JCCP of $\tilde{\mathbb{T}}_n$, truncated up to τ, is distributed as $\tilde{Z}_n$ reflected below τ. The set of levels at which births occurred on the lineage of the i-th individual, up to its coalescence with the rest of the tree, is then exactly the set of values taken by the future infimum of the i-th excursion of the JCCP under τ. First, this entails that the atoms of Σ_n are i.i.d. Second, using a time reversal argument, the distribution of this set can be read from the ascending ladder height process of $\tilde{Z}_n$. A similar reasoning for the splitting tree with mutations leads to the following facts. Consider H_n^+ the ascending ladder height process of $\tilde{Z}_n$, and put marks on its jumps in agreement with the marks on the corresponding jumps of $\tilde{Z}_n$. Note that this implies a selection of the marks that are carried by jumps of the supremum process of $\tilde{Z}_n$. Denoting by H_n^{M} the counting process of these marks, the bivariate process $(H_n^+, H_n^{\mathrm{M}})$ is a (possibly killed) bivariate subordinator which we call the marked ladder height process. The mutations on a lineage form then an inhomogeneous regenerative set, distributed as the image

by H_n^+ of the jump times of H_n^{M} under the excursion measure of $\tilde{Z}_n$ away from 0, which finally yields a simple description of the law of the (i.i.d.) atoms of Σ_n.

Convergence results Obtaining an invariance principle for a population model in a large population asymptotic requires to assume that as $n \to \infty$, the population converges in a certain sense. A classical example would be the convergence of the rescaled Bienaymé-Galton-Watson process towards the Feller diffusion [Lam67]. Now regardless of mutations, the JCCP offers a one-to-one correspondence between $\mathbb{T}_n$ and a continuous time process. Our first assumption arises then naturally as the convergence in distribution of the properly rescaled Lévy process $\tilde{Z}_n$ towards a Lévy process Z (with infinite variation, Assumption A). In particular, the lifetimes of individuals do not necessarily vanish in the limit. Besides, two different assumptions concerning the mutations are considered. The first one (B.1) falls within the classical asymptotic of rare mutations : every birth is of type 1 with a constant probability θ_n, and $\theta_n \to 0$ as $n \to \infty$. Asymptotic results in this framework are obtained in [Ber10] for the genealogical structure of alleles in a critical or subcritical Bienaymé-Galton-Watson process (however contrary to ours, they do not concern the extant population at a fixed time horizon, but the whole population). The second one (B.2) examines the case where the probability of an individual to be a mutant is correlated with her lifetime, in the sense that mutations favor longer lifetimes.
While Assumption A alone ensures the convergence in distribution of H_n^+ towards the classical ladder height process of Z, Assumptions B.1 and B.2 are designed to allow that of the marked ladder height process. Indeed, we prove in Chapter I the convergence in law of $(H_n^+, H_n^{\mathrm{M}})$ towards a (possibly killed) bivariate subordinator (H^+, H^{M}), such that H^+ is the ladder height process of Z. Note nevertheless that in this framework there is in general no convergence of the whole mutation process, namely $\tilde{Z}_n^{\mathrm{M}}$. In the case of Assumption B.1, H^+ and H^{M} are independent, and H^{M} is a Poisson process with parameter θ, which arises as the limit of the sequence θ_n after a proper rescaling. This means that the contribution to the mutations in the limit exclusively comes from individuals with vanishing lifetimes. This is no longer the case under Assumption B.2, yet additional independent marks can appear if Z has a Gaussian component. Using this convergence to deduce that of the (rescaled) law of the mutations on a lineage, the convergence of (Σ_n) to a Poisson point measure is then a straightforward consequence of the law of rare events for null arrays (see e.g. [Kal02, Th.16.18] or Corollary 0.3). Under B.1, its intensity measure is the law of the image by H^+ of an independent Poisson process with parameter θ, under the excursion measure of Z away from zero. A very similar but slightly more complicated result, involving the limiting marked ladder height process (H^+, H^{M}), is available under B.2. Besides, in the case where Z is a Brownian motion, H^+ is simply a drift, and thus the intensity measure is the law of a Poisson process killed at some independent random time, distributed as the depth of an excursion of the Brownian motion away from 0.

Outline The paper is organized as follows : Section 2 is devoted to the statement of our results, and Section 3 to their proofs. In the appendix, we give proof of some properties that are consequences of Assumption A, and which we make frequent use of throughout the paper.

2 A limit theorem for splitting trees with mutations at birth

2.1 JCCP of a marked splitting tree

Formally, a splitting tree (without mutations) is a random real tree characterized by a σ-finite measure Λ on $(0, \infty)$, satisfying $\int (1 \wedge u)\Lambda(\mathrm{d}u) < \infty$. Consider such a splitting tree, and assume first that there is extinction of the population. In [Lam10], A. Lambert considers a contour

process of this tree called JCCP (jumping chronological contour process). He establishes that the tree and its contour process are in one-to-one correspondence and characterizes the law of the latter : conditional on the first individual in the tree to have life duration x, its JCCP is distributed as a finite variation, spectrally positive Lévy process with drift -1 and Lévy measure Λ, starting at x, and killed upon hitting 0. In the case of non extinction, we then can consider the JCCP of the tree truncated up to level τ, which has the law of the Lévy process described above, starting at $x \wedge \tau$, and reflected below level τ. As noticed in Section 1, the exploration of the tree by its JCCP defines a way of labelling the individuals. In the sequel, when we label the extant individuals at level τ, we refer to that order.

Consider now a marked splitting tree $\mathbb{T}$ as defined in Section 1. We assume that the probability for a child to be a mutant can only (possibly) depend on her life span u, and if we denote by $f(u)$ this probability, where f is a function from $\mathbb{R}_+^*$ to $[0,1]$, f will be called the *mutation function* of the tree. Then $\mathbb{T}$ is characterized by its mutation function f and its lifespan measure Λ.

Then similarly as in the case without mutations, we define the JCCP of $\mathbb{T}$. First assume that there is extinction of its population. Then the JCCP of the marked tree $\mathbb{T}$ is a bivariate process (Z, Z^{M}) from $\mathbb{R}_+$ to $\mathbb{R}_+ \times \mathbb{Z}_+$, whose first coordinate Z is the JCCP of the splitting tree without marks, and whose second coordinate Z^{M} is the counting process of the mutations (see Figure 1). More precisely, for every jump time of Z (which corresponds to the encounter of a birth event in the exploration process), Z^{M} jumps (with amplitude 1) iff this birth was a 1-type birth. Hereafter we say that a jump of Z occurring at time t carries a mark (or a mutation) if $\Delta Z^{\mathrm{M}}(t) = 1$.

This bivariate process is in one-to-one correspondence with $\mathbb{T}$. Besides, conditional on the first individual to have life duration x, it is distributed as a bivariate Lévy process with drift $(-1, 0)$, and Lévy measure $\Lambda(\mathrm{d}u)\mathbb{B}_{f(u)}(\mathrm{d}q)$ (where $\mathbb{B}_r$ denotes the Bernoulli probability measure with parameter r), starting at $(x, 0)$, and killed as soon as its first coordinate hits 0. As in the non-marked case, if the assumption of extinction does not hold, the law of the JCCP of the truncated tree can be obtained from the Lévy process we just described.

2.2 Definitions and notation

2.2.1 Rescaling the population

Let $(\Lambda_n)_{n\geq 1}$ be a sequence of measures on $(\mathbb{R}_+^*, \mathcal{B}(\mathbb{R}_+^*))$ satisfying $\int (1 \wedge u)\Lambda_n(\mathrm{d}u) < \infty$ for all n, and $(f_n)_{n\geq 1}$ a sequence of continuous functions from $\mathbb{R}^+$ to $[0,1]$.

We now consider a sequence of marked splitting trees $(\mathbb{T}_n)_{n\geq 1}$ such that for all n, $\mathbb{T}_n$ has lifespan measure Λ_n, and mutation function f_n. Recalling that $\mathbb{B}_r$ denotes the Bernoulli probability measure with parameter r, we consider (Z_n, Z_n^{M}) an independent bivariate Lévy process with finite variation, Lévy measure $\Lambda_n(\mathrm{d}u)\mathbb{B}_{f_n(u)}(\mathrm{d}q)$ and drift $(-1, 0)$, and make the following assumption :

Assumption A : *There exists a sequence of positive real numbers $(d_n)_{n\geq 1}$ such that as $n \to \infty$, the process defined by*

$$\tilde{Z}_n := \Big(\frac{1}{n} Z_n(d_n t)\Big)_{t\geq 0}$$

converges in distribution to a (necessarily spectrally positive) Lévy process Z with infinite variation. We denote by Λ its Lévy measure and by b its Gaussian coefficient ($b \in \mathbb{R}_+$).

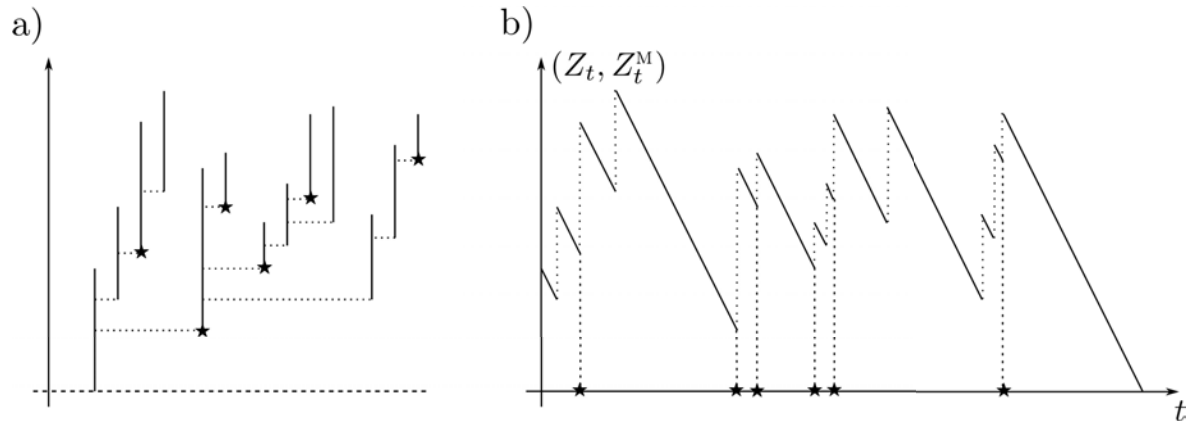

FIGURE 1 – a) A marked splitting tree : the vertical axis indicates chronological levels ; the horizontal axis has no meaning, but the horizontal lines show filiation. The marks in the tree are symbolized by stars.
b) The associated JCCP (Z, Z^{M}) : Z is the classical JCCP of the (non-marked) tree represented in a) ; the counting process of the mutations Z^{M} is not drawn as a jump process on $\mathbb{R}_+$, but is represented by the sequence of its jump times, which are symbolized by stars on the horizontal axis.

For all $n \in \mathbb{N}$ and for all $t \geq 0$, set $\tilde{Z}_n^{\mathrm{M}}(t) := Z_n^{\mathrm{M}}(d_n t)$. With an abuse of notation, the law of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ conditional on $(\tilde{Z}_n(0), \tilde{Z}_n^{\mathrm{M}}(0)) = (x, 0)$, and the law of Z conditional on $Z(0) = x$, will both be denoted by $\mathbb{P}_x$, and we write $\mathbb{P}$ for $\mathbb{P}_0$.

Denote by $\tilde{\mathbb{T}}_n$ the splitting tree obtained from $\mathbb{T}_n$ by rescaling the branch lengths by a factor $\frac{1}{n}$. The introduction of the process $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ is motivated by its fundamental role in the characterization of the law of the JCCP of $\tilde{\mathbb{T}}_n$ truncated up to level τ (see later Lemma 3.8).

Some notation : The Laplace exponents ψ_n of Z_n, $\tilde{\psi}_n$ of $\tilde{Z}_n$ and ψ of Z are defined by

$$\mathbb{E}(e^{-\lambda Z_n(t)}) = e^{t\psi_n(\lambda)},\ \mathbb{E}(e^{-\lambda \tilde{Z}_n(t)}) = e^{t\tilde{\psi}_n(\lambda)} \ \text{ and } \ \mathbb{E}(e^{-\lambda Z(t)}) = e^{t\psi(\lambda)}, \quad \lambda \geq 0.$$

We denote by $\tilde{\eta}_n$ (resp. η) the largest root of $\tilde{\psi}_n$ (resp. ψ) and by $\tilde{\phi}_n$ (resp. ϕ) the inverse of $\tilde{\psi}_n$ (resp. ψ) on $[\tilde{\eta}_n, \infty)$ (resp. $[\eta, \infty)$). We denote by $\tilde{W}_n$ (resp. W) the scale function of $\tilde{Z}_n$ (resp. Z). Finally, we denote by $\tilde{\Lambda}_n$ the Lévy measure of $\tilde{Z}_n$.

Remarks about (d_n) : Writing for $\lambda \geq 0$, $\mathbb{E}(e^{-\lambda \tilde{Z}_n(t)}) = e^{d_n t \psi_n(\lambda/n)}$, we get from the Lévy-Khintchine formula [Del13a, (2)] that $\tilde{Z}_n$ has drift $-\frac{d_n}{n}$, Lévy measure $\tilde{\Lambda}_n = d_n \Lambda_n(n\cdot)$ and Laplace exponent $\tilde{\psi}_n = d_n \psi(\cdot/n)$. In particular, this gives $\tilde{W}_n(0) = n/d_n$. We prove in the appendix that $\tilde{W}_n$ converges pointwise to W as $n \to \infty$, and besides, the assumption of infinite variation of Z ensures $W(0) = 0$. Thereby we know that necessarily $\frac{d_n}{n} \to \infty$ as $n \to \infty$.

2.2.2 Asymptotic for the mutations

In order to allow the convergence in distribution of the mutation levels on the lineages, we have to make some technical assumptions on the mutation functions f_n. Here we suggest two possible assumptions : in the first one, the probability of a child in $\mathbb{T}_n$ to be a mutant is constant, while in the second one, this probability depends on her life duration.

Assumption B.1 :

(a) *For all $n \geq 1$, for all $u \in \mathbb{R}_+$, $f_n(u) = \theta_n$, where $\theta_n \in [0,1]$.*

(b) *As $n \to \infty$, $\frac{d_n}{n}\theta_n$ converges to some nonnegative real number θ.*

Assumption B.2 : *There exists f a continuous function from $\mathbb{R}_+$ to $\mathbb{R}_+$, and $\kappa \in \mathbb{R}_+$, such that :*

(a) *the sequence $\left(u \mapsto \frac{f_n(nu)}{1 \wedge u}\right)$ converges uniformly to $u \mapsto \frac{f(u)}{1 \wedge u}$ on $\mathbb{R}_+^*$,*

(b) *$f(u)/u \to \kappa$ as $u \to 0^+$.*

Note that in B.1, necessarily $\theta_n \to 0$ as $n \to \infty$, corresponding to the classical rare mutation asymptotic. Then if we denote by f the limit of the sequence (f_n), we have $f \equiv 0$. Besides, in Assumption B.2 the choice of f_n and f is independent of $\tilde{Z}_n$ and Z.

Remark 2.1. *These two possible assumptions for the rescaling of the mutations have been chosen so that as $n \to \infty$, the marked coalescent point process converges. However this choice does not imply, despite Assumption A, the convergence of the bivariate process $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$. As pointed out in Chapter I, it is even never the case under B.2.*

2.2.3 Marked genealogical process

From now on, we consider the sequence of rescaled marked splitting trees $(\tilde{\mathbb{T}}_n)$, and condition $\tilde{\mathbb{T}}_n$ on having I_n extant individuals at level τ, where $I_n \sim \frac{d_n}{n}$ as $n \to \infty$.

We consider the space of positive point measures on $(0,\tau) \times \{0,1\}$, and endow it with the σ-field generated by the mappings $\{p_B : \xi \mapsto \xi(B),\ B \in \mathcal{B}((0,\tau)) \otimes \mathscr{P}(\{0,1\})\}$. Then we denote by $\mathcal{M}_{\mathrm{P}}$ the subset of the point measures on $(0,\tau) \times \{0,1\}$ of the form

$$\delta_{(a_m,0)} + \sum_{i=0}^{m-1} \delta_{(a_i,1)}, \text{ where } m \in \mathbb{Z}_+ \text{ and } 0 < a_0 < \ldots < a_{m-1} \leq a_m < \tau.$$

Consider a realization of $\tilde{\mathbb{T}}_n$, and label the I_n individual alive at τ from 0 to $I_n - 1$ (according to Section 2.1). Then to the i-th one we associate a simple point measure $\sigma_n^{(i)}$, with values in $(0,\tau) \times \{0,1\}$, as follows :
Consider the lineage of individual i, and assume it contains M 1-type birth events. Denote by m_0 the level where the lineage coalesces with the rest of the tree, and by $m_j,\ 1 \leq j \leq M$ the successive levels (in increasing order) where the 1-type birth events happened. Then we set

$$\sigma_n^{(i)} := \delta_{(\tau - m_0, 0)} + \sum_{1 \leq j \leq M} \delta_{(\tau - m_j, 1)}.$$

Hence the point measure $\sigma_n^{(i)}$ is in the space $\mathcal{M}_{\mathrm{P}}$, and keeps record of all the mutation events on the i-th lineage, and of the coalescence level of this lineage with the rest of the tree (see Figure 2). The quantity $\tau - m_0$ will be called the coalescence time of the lineage (the word 'time' is here to interpret as a duration). Note that in case the coalescence corresponds to a 1-type birth

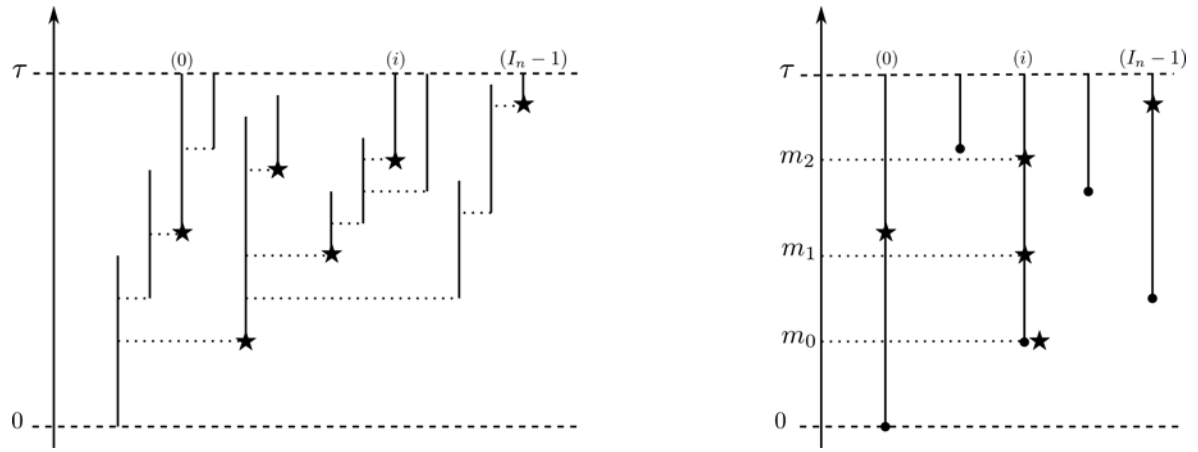

FIGURE 2 – A marked splitting tree truncated up to level τ and the associated marked coalescent point process. The 1-type birth events are symbolized by stars, and dots represent coalescence levels. In this example, the coalescence between the lineages of individuals i and $i-1$ coincides with a 1-type birth event, and we have $\sigma_n^{(i)} = \delta_{(\tau-m_0,0)} + \delta_{(\tau-m_0,1)} + \delta_{(\tau-m_1,1)} + \delta_{(\tau-m_2,1)}$.

event, we have $m_0 = m_1$.

Now for all $n \geq 1$, we define the following random point measure on $[0,1] \times \mathcal{M}_{\mathrm{P}}$:

$$\Sigma_n := \sum_{i=1}^{I_n - 1} \delta_{\{\frac{in}{d_n}, \sigma_n^{(i)}\}}.$$

The first individual (labeled 0) is on purpose not taken in account (see Remark 2.4 below). The point measure Σ_n is called the marked coalescent point process of $\tilde{\mathbb{T}}_n$. As announced in Section 1, the aim of this paper is to obtain a convergence theorem for Σ_n in a large population asymptotic.

2.3 Main results

We first introduce some notation. To begin with, we define the mapping Ψ as follows (see figure 3.a) : for all $(h, \mathbf{u} = (u_i)_{i\geq 1}, l) \in \mathbb{D}((0,\tau)) \times (\mathbb{R}_+)^{\mathbb{N}} \times \mathbb{R}_+$,

$$\Psi(h, \mathbf{u}, l) = \delta_{(h(l-),0)} + \sum_{i=1}^{j(\mathbf{u},l)} \delta_{(h(u_i),1)},$$

where

$$j : \left\{ \begin{array}{rcl} (\mathbb{R}_+)^{\mathbb{N}} \times \mathbb{R}_+ & \to & \mathbb{N} \cup \{+\infty\} \\ (\mathbf{u}, l) & \mapsto & \sup\{i \geq 1,\ u_i \leq l\} \end{array} \right.$$

The function Ψ has values in the point measures on $(0,\tau) \times \{0,1\}$, and if $j(\mathbf{u},l) < +\infty$ and $h(u_1) < \ldots < h(u_{j(\mathbf{u},l)}) \leq h(l-)$, then $\Psi(h,\mathbf{u},l)$ is in the set $\mathcal{M}_{\mathrm{P}}$.

For any càd-làg piecewise-constant function $g : \mathbb{R}_+ \to (0,\tau)$, if $(g_i)_{i\geq 1}$ denotes the sequence of its jump times (with $g_1 = 0$ in case $g(0) > 0$), we will use the notation $\Psi(h,g,l)$ instead of $\Psi(h,(g_i),l)$.

We denote by $\bar{Z}(t) := \sup_{[0,t]} Z$ the current supremum process of Z, and by $H^+ := \bar{Z} \circ L^{-1}$ the ladder height process of Z, where L is a local time at the supremum for Z, which will be specified later (see Section 3.1.2), and L^{-1} its inverse local time. We denote by T^A the first entrance time of Z in the Borel set A, and write T^x for $T^{\{x\}}$.

Finally, we denote by N' the excursion measure of Z away from zero (see Section 0.1.3), and we choose the normalization of the local time $\mathscr{L}$ according to [OP09], i.e. $\mathscr{L}$ satisfies the equality $\mathbb{E}\big(\int_{(0,\infty)} e^{-t}\mathrm{d}\mathscr{L}_t\big) = \phi'(1)$. Recall that for $\epsilon \in \mathscr{E}'$, $\chi(\epsilon)$ denotes its first entrance time into $[0,\infty)$. Define $\mathscr{E}''$ the set of all càd-làg functions ϵ with lifetime $\zeta < \infty$, such that $\epsilon(0) = \epsilon(\zeta) = 0$ and $\epsilon(x) > 0$ for all $0 < x < \zeta$. Then we define a measure N'' on $\mathscr{E}'' \times \{0,1\}$, endowed with the topology induced by the Skorokhod topology, as follows (see Figure 3.b). First, let $\tilde{N}'$ denote the pushforward measure of N' by the mapping

$$\left\{ \begin{array}{ccc} \mathscr{E}' & \longrightarrow & \mathscr{E}'' \times \mathbb{R}_+ \\ \epsilon & \longmapsto & ((-\epsilon((\chi - t)-))_{0\le t<\chi}, \Delta\epsilon(\chi)) \end{array} \right. .$$

Then for any $(\epsilon, \epsilon^{\mathrm{M}}) \in \mathscr{E}'' \times \{0,1\}$, N'' is defined by

$$N''((\epsilon,\epsilon^{\mathrm{M}}) \in \mathrm{d}E \times \mathrm{d}q) := \int_{[0,\infty)} \tilde{N}'(\mathrm{d}E \times \mathrm{d}x)\mathbb{B}_{f(x)}(\mathrm{d}q).$$

Note that

$$N''((\epsilon,\epsilon^{\mathrm{M}}) \in \mathrm{d}E \times \{0,1\}) = \tilde{N}'(\mathrm{d}E \times \mathbb{R}_+),$$

and that in the case where Z does not drift to $+\infty$, the excursions of Z have finite lifetime, so that from a time reversal argument we have for any measurable set E of $\mathscr{E}''$, $N''(E \times \{0,1\}) = N'(\epsilon|_{[\chi,\zeta)} \in E)$, where $\epsilon|_{[\chi,\zeta)}$ denotes the restriction of ϵ to the interval $[\chi,\zeta)$.

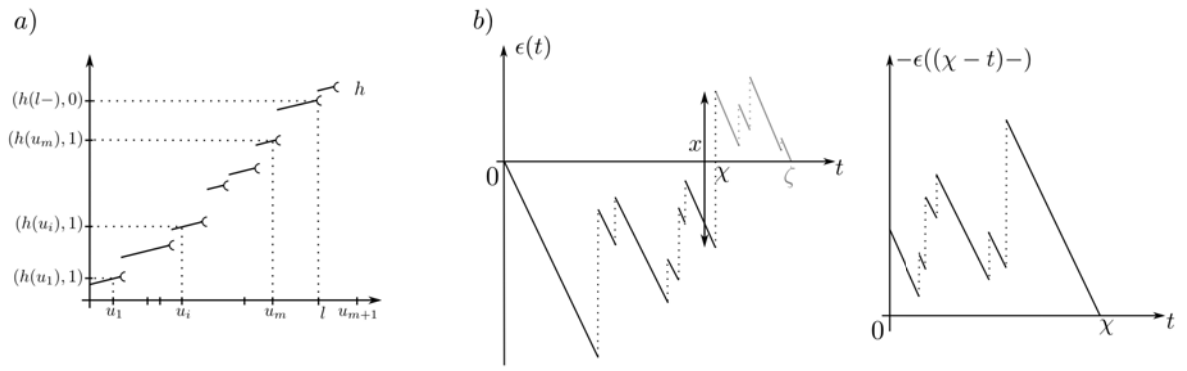

FIGURE 3 – a) A graphical representation of a triplet $(h, \mathbf{u} = (u_i)_{i\ge1}, l) \in \mathbb{D}((0,\tau)) \times (\mathbb{R}_+)^{\mathbb{N}} \times \mathbb{R}_+$. In this example, $j(\mathbf{u}, l) = m$ and $\Psi(h, \mathbf{u}, l) = \delta_{(h(l-),0)} + \sum_{i=1}^m \delta_{(h(u_i),1)}$.
b) Left panel : A representation (in finite variation) of an excursion $\epsilon \in \mathscr{E}'$, with finite lifetime ζ, such that $\Delta\epsilon(\chi) = x$. Right panel : The corresponding reversed excursion on $[0,\chi)$: $(-\epsilon(\chi - t)-))_{0\le t<\chi}$ (which belongs to $\mathscr{E}''$).

Results under Assumption B.1

In this paragraph we suppose that Assumptions A and B.1 are satisfied.

Theorem 2.2. *Consider an independent Poisson process Θ with parameter θ. We introduce σ, a random element of $\mathcal{M}_{\mathrm{P}}$, defined on $\{T^0 < \infty\}$ by*

$$\sigma = \Psi(H^+, \Theta, L(T^0)).$$

Then the sequence (Σ_n) converges in distribution towards a Poisson point measure Σ on $[0,1] \times \mathcal{M}_{\mathrm{P}}$ with intensity measure $\mathrm{Leb} \otimes \Pi_1$, where Π_1 is a measure on $\mathcal{M}_{\mathrm{P}}$ defined by

$$\Pi_1 = N''(\sigma \in \cdot,\ \sup \epsilon < \tau).$$

Remark 2.3. *Denote by $B_{\geq m} := \{\sigma \in \mathcal{M}_{\mathrm{P}},\ \sigma((0,\tau) \times \{1\}) \geq m\}$ the set of point measures of $\mathcal{M}_{\mathrm{P}}$ having at least m points with second coordinate 1 in the interval $(0,\tau)$, which can be interpreted here as the presence of at least m mutations on a lineage. Then the measure $\Pi_1(B_{\geq 1})$ is not necessarily finite (see Example 1).*

Remark 2.4. *Note that we excluded in Σ_n the first lineage $\sigma_n^{(0)}$, for which without additional assumption, we cannot easily get a similar result as for the other lineages. However, if we assume that the lifetime of the first individual in $\tilde{\mathbb{T}}_n$ converges as $n \to \infty$ towards some value greater than τ, we can adapt Theorem 2.2. The limiting object is then obtained by adding to Σ a Dirac mass on $(0, \delta_{(\tau,0)})$.*

Remark 2.5. ***Conditioning $\tilde{\mathbb{T}}_n$ on survival at level τ***
We obtain a similar result if, instead of conditioning $\tilde{\mathbb{T}}_n$ on having I_n extant individuals at level τ, we condition it on survival at level τ. Indeed, if we denote by $\tilde{\Xi}_n(\tau)$ the number of extant individuals in $\tilde{\mathbb{T}}_n$ at level τ, we know that conditional on $\tilde{\Xi}_n(\tau) \geq 1$, $\tilde{\Xi}_n(\tau)$ follows a geometric distribution with parameter $\frac{n}{d_n \tilde{W}_n(\tau)}$ (see [Lam10, prop.5.6]). Then thanks to the pointwise convergence of $\tilde{W}_n$ towards W (see Proposition 3.1), we get that $\frac{n}{d_n}\tilde{\Xi}_n(\tau)$ converges in distribution towards an exponential variable with parameter $\frac{1}{W(\tau)}$.
Then the sequence (Σ_n) converges in law to a Poisson point measure on $[0, \boldsymbol{e}] \times \mathcal{M}_{\mathrm{P}}$ with intensity $\mathrm{Leb} \otimes \Pi_1$, where $\boldsymbol{e}$ is an independent exponential variable with parameter $\frac{1}{W(\tau)}$.

Assume now that Z has **no Gaussian component**, and let Θ be as in Theorem 2.2. Then using Proposition 0.7 we get

$$\Pi_1 = \int_{(0,\tau)} \mathrm{d}x\, \bar{\Lambda}(x)\, \mathbb{P}_x(\sigma \in \cdot,\ T^0 < T^{(\tau,\infty)}),$$

where σ is defined in Theorem 2.2 and $\bar{\Lambda}(x) = \Lambda((x,\infty))$ for all $x > 0$. Hence in the limit, the mutations appearing on a lineage are distributed according to a point measure σ, where σ is distributed as the image of the jump times of an independent Poisson process with parameter θ, by the ladder height process of Z conditioned on $T^0 < T^{(\tau,\infty)}$, and starting at the opposite of the undershoot of an excursion with depth smaller than τ.

Finally, the following proposition expresses the law of σ under $\mathbb{P}_x(\cdot \cap \{T^0 < T^{(\tau,\infty)}\})$ in terms of the image of an independent Poisson process by an inhomogeneous killed subordinator.

Proposition 2.6. *Let Θ and σ be as in Theorem 2.2. For all $x \in (0,\tau)$,*

$$\mathbb{P}_x(\sigma \in \cdot,\ T^0 < T^{(\tau,\infty)}) = \mathbb{P}_x(\sigma^{\mathrm{K}} \in \cdot),$$

where

$$\sigma^{\mathrm{K}} := \Psi(H^{\mathrm{K}}, \Theta, \ell),$$

with H^{K} a killed inhomogeneous subordinator with drift $\frac{b^2}{2}$ and jump measure μ^{K}, defined for all $a \in (0,\tau)$ and $u \in (0,\tau-a)\times\{+\infty\}$ by :

$$\mu^{\mathrm{K}}(a,\mathrm{d}u) := \frac{1}{W(a)}\delta_{+\infty}(\mathrm{d}u) + \int_{(0,a)} \mathrm{d}x\, \Lambda(x+\mathrm{d}u)\, \frac{W(a-x)W(\tau-a-u)}{W(a)W(\tau)},$$

and $\ell := \inf\{t \geq 0,\ H^{\mathrm{K}}(t) = +\infty\}$ the killing time of H^{K}.

Results under Assumption B.2

We suppose now that Assumptions A and B.2 are satisfied. We establish in this case some very similar results as under B.1, but in a slightly more complicated version. Indeed, Assumption B.1 ensures the independence of H^+ with a certain process we define later (namely the subordinator H^{M} that appears in the following statement), while in case B.2 these two subordinators are no longer independent.

Theorem 2.7. *There exists a process H^{M}, starting at 0 under N'', such that (H^+, H^{M}) is a (possibly killed) bivariate subordinator, and such that (Σ_n) converges in distribution towards a Poisson point measure Σ on $[0,1]\times\mathcal{M}_{\mathrm{P}}$ with intensity measure $\mathrm{Leb}\otimes\Pi_2$, where*

$$\Pi_2 = N''(\,\Psi(H^+, \epsilon^{\mathrm{M}} + H^{\mathrm{M}}, L(T^0)) \in \cdot,\ \sup\epsilon < \tau).$$

The processes H^+ and H^{M} are not independent unless Z is a Brownian motion with drift, and the law of (H^+, H^{M}) is explicitly characterized in Theorem 3.3.

Note that Remarks 2.4 and 2.5 are still relevant in case B.2.

Remark 2.8. *If the limiting process Z is a Brownian motion with drift, H^+ is a deterministic drift and hence H^+ and H^{M} are automatically independent. Hence in this case, Theorem 2.2 remains valid under Assumption B.2.*

Similarly as under B.1, if Z has no Gaussian component we can reexpress the measure Π_2 as follows :

$$\Pi_2 = \int_{(0,\tau)\times\{0,1\}} \mathrm{d}x \int_{(x,\infty)} \Lambda(\mathrm{d}u)\, \mathbb{B}_{f(u)}(\mathrm{d}q)\, \mathbb{P}_x(\sigma_q \in \cdot,\ T^0 < T^{(\tau,\infty)}),$$

where for $q \in \{0,1\}$, $\sigma_q = \Psi(H^+, q + H^{\mathrm{M}}, L(T_0))$.

Furthermore, as in Proposition 2.6, we have for all $x \in (0,\tau)$, $q \in \{0,1\}$

$$\mathbb{P}_x(\sigma_q \in \cdot,\ T^0 < T^{(\tau,\infty)}) = \mathbb{P}_x(\sigma_q^{\mathrm{K}} \in \cdot),$$

where

$$\sigma_q^{\mathrm{K}} := \Psi(H^{\mathrm{K}}, q + H^{\mathrm{K,M}}, \ell),$$

with $(H^{\mathrm{K}}, H^{\mathrm{K,M}})$ a bivariate killed inhomogeneous subordinator, starting at $(x,0)$ under $\mathbb{P}_x$, with drift $(\frac{b^2}{2}, 0)$ and jump measure μ^{K}, defined for all $a \in (0,\tau)$, $u \in (0,\tau-a)\times\{+\infty\}$ and $q \in \{0,1\}$ by :

$$\mu^{\mathrm{K}}(a,\mathrm{d}u,\mathrm{d}q) := \frac{1}{W(a)}\delta_{(+\infty,0)}(\mathrm{d}u,\mathrm{d}q) + \int_{(0,a)} \mathrm{d}x\, \Lambda(x+\mathrm{d}u)\, \mathbb{B}_{f(x+u)}(\mathrm{d}q)\, \frac{W(a-x)W(\tau-a-u)}{W(a)W(\tau)},$$

and $\ell := \inf\{t \geq 0,\ H^{\mathrm{K}}(t) = +\infty\}$ the killing time of H^{K}.

We close this section by giving some explicit calculations in the cases where the limiting process Z is either the standard Brownian motion, or an α-stable Lévy process ($\alpha \in (1,2)$).

Example 1 : The Brownian case
Consider the case where the population of $\mathbb{T}_n$ have exponential life spans with mean 1. Then an appropriate rescaling of the JCCP of $\mathbb{T}_n$ leads in the limit to the standard Brownian motion.

We set :

$$\Lambda_n(\mathrm{d}r) = e^{-r}\mathbb{1}_{r\geq 0}\mathrm{d}r \quad \text{and} \quad d_n = \frac{n^2}{2}.$$

Then, Assumption A is satisfied : for all $\lambda \geq 0$, we have $\tilde{\psi}_n(\lambda) = \frac{n}{\lambda+n}\frac{\lambda^2}{2}$, which converges to $\psi(\lambda) = \frac{\lambda^2}{2}$ as $n \to \infty$, and this implies the convergence in $\mathbb{D}(\mathbb{R})$ of $\tilde{Z}_n$ towards the standard Brownian motion (see [JS87, Th.VII.3.4]). Moreover, if we assume $\theta_n = \frac{\beta}{n}$ for some $\beta \in [0,1]$, Assumption B.1 holds with $\theta = \frac{\beta}{2}$.

The genealogical structure of this process (without mutations) and its asymptotic behaviour are studied by L. Popovic in [Pop04], and in particular, results taking into account a β-sampling of extinct individuals (each individual in the genealogy is recorded with a probability β) are provided. The following results are presented as a consequence of Theorem 2.2 but can also be derived from [Pop04], since β-sampling can be directly interpreted as recording 1-type birth events in the genealogy.

The distribution of Σ is completely explicit. We know that $W(x) = 2x$, and $H^+(t) = \frac{t}{2}$ a.s. for all $t \geq 0$. Note that the image by H^+ of a Poisson process with parameter θ is itself a Poisson process, with parameter 2θ. As a consequence, if we denote by $((a_0,0),(a_1,1)...,(a_j,1))$ the ranked sequence of the atoms of the measure σ appearing in Theorem 2.2, under $N''(\,\cdot \cap \sup\epsilon \in (0,\tau))$, conditional on a_0, $(a_1,...,a_j)$ is distributed as the sequence of jump times of a Poisson process with parameter β, restricted to $(0,a_0)$.

Besides, from the criticality of Brownian motion, we have $N''(\,\cdot \times \{0,1\}) = N'(\epsilon|_{[\chi,\zeta)} \in \cdot\,)$, and since an excursion of Brownian motion away from 0 is such that $\chi = 0$ or $\chi = \zeta$,

$$N''(\sigma \in \cdot,\ \sup\epsilon < \tau) = N'(\sigma \in \cdot,\ \sup\epsilon \in (0,\tau)).$$

Finally, we have

$$N'(H^+(L(T^0-)) \in \mathrm{d}h,\ \sup\epsilon \in (0,\tau)) = N'(\sup\epsilon \in \mathrm{d}h,\ \sup\epsilon \in (0,\tau)) = \frac{\mathrm{d}h}{2h^2}\mathbb{1}_{0<h<\tau}.$$

The measure Π_1 can then be expressed as follows :

$$\Pi_1 = \int_0^\tau \frac{\mathrm{d}h}{2h^2}\int_{\mathcal{M}} \pi_{\beta,h}(\mathrm{d}\Theta)\mathbb{1}_{\{\delta_{(h,0)}+\sum_{i\in I}\delta_{(\Theta_i,1)}\in\cdot\,\}},$$

where $\mathcal{M}$ denotes the space of point measures on $\mathbb{R}_+$, $\pi_{\beta,h}$ is the law of a Poisson process with parameter β restricted to the interval $(0,h)$, and for any $\Theta \in \mathcal{M}$, $(\Theta_i)_{i\in I}$ denotes the sequence of jump times of Θ.

In other words, in the limit the mutations on a lineage are distributed as an independent Poisson process with parameter β, stopped at an independent random time distributed as the depth of an excursion away from 0, with depth lower than τ. Note furthermore that simple calculations lead to $\Pi_1(B_{\geq 1}) = \infty$ and $\Pi(B_{\geq 2}) < \infty$ (using the notation introduced in Remark 2.3) : the number of lineages carrying at least one mutation (resp. two mutations) is a.s. infinite (resp. finite).

Moreover, contrary to what is announced in Remark 2.4, the loss of memory of the exponential distribution ensures here that there is no need to add extra assumptions to extend the result to the first lineage. In this case, the limiting object is then obtained by adding to Σ a Dirac mass on $(0, \sigma_\tau + \delta_{(\tau,0)})$ where σ_τ is an independent Poisson process on $[0, \tau)$ with parameter β.

Finally, according to Remark 2.8, these results are still valid for any choice of a sequence of functions (f_n) and of a real number κ satisfying B.2, replacing β by 2κ.

Example 2 : The stable case
Fix $\alpha \in (1, 2)$ and set :

$$\Lambda_n(\mathrm{d}r) = -\frac{r^{-\alpha-1}}{\Gamma(-\alpha)}\mathbb{1}_{r>1}\mathrm{d}r \quad \text{and} \quad d_n = n^\alpha,$$

then we have for all $\lambda \geq 0$, $\tilde{\psi}_n(\lambda) \to \lambda^\alpha$ which is the Laplace exponent of an α-stable spectrally positive Lévy process and Assumption A is satisfied. If we now set $\theta_n := \beta/n^{\alpha-1}$ for some $\beta \in [0, 1]$, Assumption B.1 holds with $\theta = \beta$.

In this case we are able to characterize explicitly the inhomogeneous killed subordinator H^{κ} defined in Proposition 2.6. Indeed, we know that Z has no Gaussian component, $\Lambda(\mathrm{d}z) = -\frac{z^{-\alpha-1}}{\Gamma(-\alpha)}\mathrm{d}z$, and $W(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}$. Hence H^{κ} has no drift and for all $a \in (0, \tau)$, $u \in (0, \tau - a) \times \{+\infty\}$, a simple calculation leads to

$$\mu^{\kappa}(a, \mathrm{d}u) = -\frac{u^{-\alpha-1}}{\Gamma(-\alpha)}\frac{au}{u+a}\left(\frac{\tau - a - u}{\tau}\right)^{\alpha-1}\mathrm{d}u + \frac{a^{\alpha-1}}{\Gamma(\alpha)}\delta_{+\infty}(\mathrm{d}u).$$

3 Proofs of statements

Proving our theorems first requires to give some preliminary results (Section 3.1), and in particular, the introduction of the marked ladder height process of $(\tilde{Z}_n, \tilde{Z}^{\mathrm{M}}_n)$ we described in the Introduction. The definition of this process and the convergence results we obtained in Chapter I are reviewed in Section 3.1.2. Then Section 3.2 is devoted to the proof of the results stated in Section 2.3, relegating to Section 3.3 the proof of some technical result of convergence.

3.1 Preliminary results

3.1.1 Consequences of Assumption A

We state here some direct consequences of the convergence of $\tilde{Z}_n$ towards Z, which we prove in the appendix. Denote by T^A_n the first entrance time of $\tilde{Z}_n$ in the Borel set A, and write T^x_n for $T^{\{x\}}_n$. Recall that similar notation has been introduced in Section 2.3 for the limiting process Z. Then Assumption A leads to :

Proposition 3.1. (i) *For all $x, y > 0$, under $\mathbb{P}_0$, T^{-x}_n (resp. $T^{(y,\infty)}_n$) converges in distribution to T^{-x} (resp. $T^{(y,\infty)}$) as $n \to \infty$.*

(ii) *As $n \to \infty$, $\tilde{\phi}_n \to \phi$ uniformly on every compact set of $\mathbb{R}_+$, and in particular $\tilde{\eta}_n \to \eta$.*

(iii) *As $n \to \infty$, $\tilde{W}_n \to W$ uniformly on $\mathbb{R}_+$, and $\tilde{W}'_n \to W'$ uniformly on every compact set of $\mathbb{R}^*_+$.*

Remark 3.2. *According to the remark after Lemma 8.2, and Exercise 8.4 in [Kyp06], in the infinite variation case the scale function of a Lévy process is differentiable on $\mathbb{R}_+^*$ with continuous derivative, and in the finite variation case, it has left and right derivatives on $\mathbb{R}_+^*$.*

3.1.2 Convergence of the marked ladder height process

In this section we define the marked ladder height process of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$, and recall the convergence results obtained for this process in Chapter I.

Local times at the supremum

We first need to specify local times at the supremum for the processes $\tilde{Z}_n$ and Z. We denote by $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ (resp. $\mathcal{F}_n = (\mathcal{F}_{n,t})_{t\geq 0}$) the natural filtration associated to Z (resp. $\tilde{Z}_n$), that is for all $t \geq 0$,

$$\mathcal{F}_t = \sigma\{Z_s,\ s \leq t\} \text{ (resp. } \mathcal{F}_{n,t} = \sigma\{\tilde{Z}_n(s),\ s \leq t\})$$

For all $n \geq 1$, let $(\tau_{n,i})_{i\geq 0}$ be a sequence of i.i.d. random exponential variables, independent of $(\tilde{Z}_n)_{n\geq 1}$, with parameter $\alpha_n := \frac{d_n}{n}$. Then, according to Section 0.1.3, we define for $\tilde{Z}_n$ a local time at the supremum as follows :

$$L_n(t) := \sum_{i=0}^{\imath_n(t)} \tau_{n,i},$$

where $\imath_n(t)$ represents the number of jumps of the supremum until time t. We denote by L_n^{-1} the right-continuous inverse of L_n, and replace the filtration $\mathcal{F}_{n,t}$ with $\mathcal{F}_{n,t} \vee \sigma(L_n(s),\ s \leq t)$, so that L_n (resp. L_n^{-1}) is adapted to $(\mathcal{F}_{n,t})$ (resp. to $(\mathcal{F}_{n,L_n^{-1}(t)})$).

We introduce the local time at the supremum L for the infinite variation Lévy process Z : it is defined up to a multiplicative constant, and we require that

$$\mathbb{E}\left(\int_{(0,\infty)} e^{-t}\mathrm{d}L_t\right) = \phi(1), \tag{1}$$

so that L is uniquely determined. Finally, we denote by L^{-1} its inverse.

Excursions and mutations

From now on, we assume (unless otherwise specified) that $\tilde{Z}_n^{\mathrm{M}}(0) = 0$ a.s. We denote by $(t, e_{n,t})_{t\geq 0}$ the excursion process of $\tilde{Z}_n$ formed by the excursions from its past supremum, and N_n its excursion measure, as defined in Section 0.1.3.
We define for all $t \in [0, L_n(\infty))$

$$\xi_n := \begin{cases} (t, e_{n,t}(\zeta), \Delta\tilde{Z}_n^{\mathrm{M}}(L_n^{-1}(t)))_{t\geq 0} & \text{if } L_n^{-1}(t-) < L_n^{-1}(t) \\ \partial & \text{else} \end{cases},$$

where ∂ is an additional isolated point, and $e_{n,t}(\zeta)$ stands for $e_{n,t}(\zeta(e_{n,t}))$.
Here the fourth coordinate $\Delta\tilde{Z}_n^{\mathrm{M}}(L_n^{-1}(t))$ is 1 or 0 whether or not the jump of $\tilde{Z}_n$ at the right end point of the excursion interval indexed by t is marked. Note that the set $\{L_n^{-1}(t)\}_{t\geq 0}$ of these right end points is exactly the set of record times of $\tilde{Z}_n$.

Marked ladder height process
Then according to Chapter I, for $n \geq 1$, we define the marked ladder height process $H_n = (H_n^+, H_n^{\mathrm{M}})$ of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$ as the (possibly killed) bivariate subordinator with no drift and whose jump point process is a.s. equal to ξ_n. Moreover, according to Proposition 3.2 in Chapter I, H_n has Lévy measure

$$\mu_n(\mathrm{d}y, \mathrm{d}q) := \int_0^\infty \mathrm{d}x\ e^{-\tilde{\eta}_n x}\ \tilde{\Lambda}_n(x + \mathrm{d}y)\ \mathbb{B}_{f_n(n(x+y))}(\mathrm{d}q), \tag{2}$$

and is killed at rate $k_n = \frac{1}{\tilde{W}_n(\infty)}$ if $\tilde{Z}_n$ is subcritical.

Note that H_n^+ is in fact the ladder height process of $\tilde{Z}_n$, i.e. for all $t \geq 0$, $H_n^+(t) = \bar{\tilde{Z}}_n(L_n^{-1}(t))$ a.s., where $\bar{\tilde{Z}}_n(t)$ denotes the current supremum of $\tilde{Z}_n$ at time t. The jumps of H_n^{M} correspond, in the local time scale, to the marks occurring at record times of $\tilde{Z}_n$. Moreover, H_n^{M} is a Poisson process with parameter $\lambda_n := \mu_n(\mathbb{R}_+^* \times \{1\})$, so that the random time

$$\mathrm{e}_n := \inf\{t \geq 0,\ H_n^{\mathrm{M}}(t) = 1\} \tag{3}$$

follows on $\{\mathrm{e}_n < L_n(\infty)\}$ an exponential distribution with parameter λ_n.

Convergence theorem for the marked ladder height process We define

$$\mu(\mathrm{d}u, \mathrm{d}q) := \int_0^\infty \mathrm{d}x\ e^{-\eta x}\ \Lambda(x + \mathrm{d}u)\ \mathbb{B}_{f(x+u)}(\mathrm{d}q),$$

and

$$\mu^+(\mathrm{d}u) := \mu(\mathrm{d}u, \{0, 1\}) = \int_0^\infty \mathrm{d}x\ e^{-\eta x}\ \Lambda(x + \mathrm{d}u).$$

We recall here Theorem 4.1 of Chapter I :

Theorem 3.3. *Under Assumption B.1, if Z does not drift to $-\infty$, the sequence of bivariate subordinators $H_n = (H_n^+, H_n^{\mathrm{M}})$ converges weakly in law to a subordinator $H := (H^+, H^{\mathrm{M}})$, where H^+ and H^{M} are independent, H^+ is a subordinator with drift $\frac{b^2}{2}$ and Lévy measure μ^+, and H^{M} is a Poisson process with parameter θ. In the case Z drifts to $-\infty$, the same statement holds but H is killed at rate $k := \frac{1}{W(\infty)}$ and the independence between H^+ and H^{M} holds only conditional on their common lifetime.*

Under Assumption B.2, the sequence of bivariate subordinators $H_n = (H_n^+, H_n^{\mathrm{M}})$ converges weakly in law to a subordinator $H := (H^+, H^{\mathrm{M}})$, which is killed at rate k if Z drifts to $-\infty$. Moreover, H has drift $(\frac{b^2}{2}, 0)$ and Lévy measure

$$\mu(\mathrm{d}u, \mathrm{d}q) + \rho\delta_0(\mathrm{d}u)\delta_1(\mathrm{d}q),$$

where $\rho := \kappa b^2$.

In particular, under Assumption B.2, if Z has no Gaussian component, the limiting marked ladder height process is a pure jump bivariate subordinator with Lévy measure μ. If Z has a Gaussian component, the fact that the « small jumps » of $\tilde{Z}_n$ generate the Gaussian part in the limit results in a drift for H^+, and possibly additional independent marks that happen with constant rate in time, as under Assumption B.1. This rate is proportional to the Gaussian coefficient (provided that $\kappa \neq 0$). Besides, note that as expected, H^+ is distributed as the classical ladder height process of Z.

An easy adaptation of the proof of this theorem yields

Theorem 3.4. *Let H_n^* be a driftless subordinator on $\mathbb{R}_+$ with Lévy measure*

$$\mu_n^*(\mathrm{d}u) := \int_{(0,\infty)} \mathrm{d}x\, e^{-\tilde{\eta}_n x}\, \tilde{\Lambda}_n(x+\mathrm{d}u)\,(1-f_n(n(x+u))).$$

Then H_n^* converges in distribution to a subordinator H^* with drift $\frac{b^2}{2}$ and Lévy measure

$$\mu^*(\mathrm{d}u) = \int_{(0,\infty)} \mathrm{d}x\, e^{-\eta x}\, \Lambda(x+\mathrm{d}u)\,(1-f(x+u)).$$

We denote by ψ_n^* and ψ^* the respective Laplace exponents of H_n^* and H^*.

Remark 3.5. *Under Assumption B.1, this theorem is not of interest, since H_n^* (resp. H^*) is equal in law to H_n^+ (resp. H^+), and so the result is given by Theorem 3.3 with $f \equiv 0$.*

Finally, we recall Theorem 5.1 of Chapter I :

Theorem 3.6. *The following convergence in distribution holds in $\mathbb{D}(\mathbb{R})^4$ as $n \to \infty$:*

$$(\tilde{Z}_n, L_n, H_n^+, H_n^{\mathrm{M}}) \Rightarrow (Z, L, H^+, H^{\mathrm{M}}).$$

3.2 Proof of main results

The proof of Theorems 2.2 and 2.7 is organized in four subsections. In the first one we describe the distribution of the point measures $\sigma_n^{(i)}$ from a family of Markov chains. More precisely, we show that these point measures are i.i.d., and that for any $\varepsilon \in (0,\tau)$, their restriction to $[\varepsilon,\tau)\times\{0,1\}$ has the law of a point measure whose set of atoms forms a Markov chain $M_{n,\varepsilon}$, killed at some first entrance time. The second one deals with the construction of the limiting Markov chain M_ε, and then with the proof of theorems themselves, in which we make use of the convergence in distribution of $(M_{n,\varepsilon})_n$. The proof of the latter convergence is quite long and is gathered in the last two subsections.

3.2.1 Distribution of the point measures $\sigma_n^{(i)}$

From the article [Lam10] of A. Lambert, we know that there is a one-to-one correspondence between a splitting tree and its JCCP. In particular, properties linked to the lineage of the i-th extant individual at level τ are read from the i-th excursion under level τ of the truncated JCCP. Then using the invariance by time reversal of such excursions, and making use of the strong Markov property, we obtain the following proposition. Recall that we conditioned $\tilde{\mathbb{T}}_n$ on having I_n extant individuals alive at τ.

Proposition 3.7. *Fix $\varepsilon \in (0,\tau)$, $n \geq 1$, and let $\sigma_{n,\varepsilon}^{(i)}$ denote the trace measure of $\sigma_n^{(i)}$ on $[\varepsilon,\tau)\times\{0,1\}$. Then we have :*

(i) The random measures $(\sigma_{n,\varepsilon}^{(i)})_{1\leq i<I_n}$ are i.i.d.

(ii) There exists a Markov chain $M_{n,\varepsilon}$ with values in $[\varepsilon,\tau)\times\{0,1\}$ such that with probability $1-p_{n,\varepsilon}$, $\sigma_{n,\varepsilon}^{(1)}([\varepsilon,\tau)\times\{0,1\}) = 0$, and with probability $p_{n,\varepsilon}$, $\sigma_{n,\varepsilon}^{(1)}$ is distributed as

$$\sum_{k=0}^{K_{n,\varepsilon}} \delta_{M_{n,\varepsilon}(k)},$$

where $p_{n,\varepsilon} := \frac{n}{d_n}\frac{\frac{1}{\tilde{W}_n(\varepsilon)}-\frac{1}{\tilde{W}_n(\tau)}}{1-\frac{\tilde{W}_n(0)}{\tilde{W}_n(\tau)}}$, and $K_{n,\varepsilon} := \inf\{k\geq 0,\ M_{n,\varepsilon}^2(k)=0\}$ ($M_{n,\varepsilon}^i$, $i\in\{1,2\}$, denoting the i-th coordinate of $M_{n,\varepsilon}$).

The probability $p_{n,\varepsilon}$ has in fact to be interpreted as follows : we have

$$\mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}) = \mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)}) - \mathbb{P}_0(T_n^{-\tau} < T_n^{(0,\infty)}) = \frac{\tilde{W}_n(0)}{\tilde{W}_n(\varepsilon)} - \frac{\tilde{W}_n(0)}{\tilde{W}_n(\tau)},$$

and hence

$$p_n = \mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)} \,|\, T_n^{(0,\infty)} < T_n^{-\tau}).$$

Construction of $M_{n,\varepsilon}$

We construct below the Markov chain $M_{n,\varepsilon}$ appearing in Proposition 3.7, and which will converge in distribution towards the Markov chain M_ε that appears in Theorem 2.7 (the proof of the latter point is the purpose of Sections 3.3.1 and 3.3.2).

Recall that we defined in Section 3.1.2 (formula (3)) the random variable $\mathrm{e}_n = \inf\{t \geq 0,\ H_n^{\mathrm{M}}(t) = 1\}$. We set for all $n \geq 1$, $x > 0$ and $u \geq 0$:

$$\nu_n^{\mathrm{M}}(x, \mathrm{d}u) := \mathbb{P}_0\big(H_n^+(\mathrm{e}_n) \in \mathrm{d}u, L_n^{-1}(\mathrm{e}_n) < T_n^{-x} \,|\, T_n^{-x} < T_n^{(\tau-x,\infty)}\big) \tag{4}$$

$$\nu_n^{\mathrm{D}}(x, \mathrm{d}u) := \mathbb{P}_0\big(\tilde{\bar{Z}}_n(T_n^{-x}) \in \mathrm{d}u, L_n^{-1}(\mathrm{e}_n) \geq T_n^{-x} \,|\, T_n^{-x} < T_n^{(\tau-x,\infty)}\big), \tag{5}$$

where the letters M and D stand respectively for « mutation » and « death ».

We want to initialize the Markov chain $M_{n,\varepsilon}$ at the first 1-type birth event that occurs below level $\tau - \varepsilon$, when following the lineage backward in time. Then, conditional on $\tilde{Z}_n(0) = 0$ and $T_n^{-\varepsilon} < T_n^{(0,\infty)} < \infty$, we set

$$S_n := \sup\{t \leq T_n^{(0,\infty)},\ \tilde{Z}_n(t) < -\varepsilon\} \ \text{ and } (\Upsilon_n, \Upsilon_n^{\mathrm{M}}) := (-\tilde{Z}_n(S_n-), \Delta\tilde{Z}_n^{\mathrm{M}}(S_n)).$$

Thereby if we consider an excursion of $\tilde{Z}_n$ away from 0, Υ_n is the value of $\tilde{Z}_n$ before its last jump over level $-\varepsilon$ (see Figure 4), and Υ_n^{M} is the mark carried by this jump.

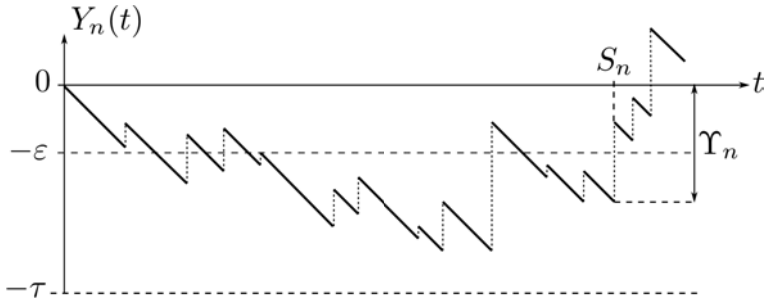

FIGURE 4 – An excursion of $\tilde{Z}_n$ under level 0 and the random variables S_n and Υ_n.

Finally we define, for all $(u, q) \in (\varepsilon, \tau) \times \{0, 1\}$,

$$\nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}u, \mathrm{d}q) := \frac{1}{p'_{n,\varepsilon}} \mathbb{P}_0(\Upsilon_n \in \mathrm{d}u,\ \Upsilon_n^{\mathrm{M}} \in \mathrm{d}q,\ T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}),$$

where $p'_{n,\varepsilon} := \frac{n}{d_n}\left(\frac{1}{\tilde{W}_n(\varepsilon)} - \frac{1}{\tilde{W}_n(\tau)}\right) = p_{n,\varepsilon}\left(1 - \frac{\tilde{W}_n(0)}{\tilde{W}_n(\tau)}\right)$ is in fact equal to $\mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau})$, and is therefore a normalizing constant such that $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ is a probability measure.

Then we consider the Markov chain $M_{n,\varepsilon} = (M_{n,\varepsilon}(k))_{k\in\mathbb{Z}_+}$ with values in $[\varepsilon, \tau) \times \{0, 1\}$, defined by :

- For all $k \in \mathbb{Z}_+$, for all $u \geq 0$, conditional on $M_{n,\varepsilon}(k) = (x, 1)$,

$$\begin{cases} M_{n,\varepsilon}(k+1) \in (x + \mathrm{d}u) \times \{1\} & \text{with probability } \nu_n^{\mathrm{M}}(x, \mathrm{d}u) \\ M_{n,\varepsilon}(k+1) \in (x + \mathrm{d}u) \times \{0\} & \text{with probability } \nu_n^{\mathrm{D}}(x, \mathrm{d}u). \end{cases}$$

- For all $k \in \mathbb{Z}_+$, conditional on $M_{n,\varepsilon}(k) = (x, 0)$, $M_{n,\varepsilon}(k+1) = (x, 0)$ a.s.

- For all $u \in [\varepsilon, \tau)$,

$$\begin{cases} \mathbb{P}(M_{n,\varepsilon}(0) \in \mathrm{d}u \times \{1\}) = \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}u \times \{1\}) + \int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x \times \{0\}) \nu_n^{\mathrm{M}}(x, \mathrm{d}u - x) \\ \mathbb{P}(M_{n,\varepsilon}(0) \in \mathrm{d}u \times \{0\}) = \int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x, \{0\}) \nu_n^{\mathrm{D}}(x, \mathrm{d}u - x) \end{cases}$$

Recall that $K_{n,\varepsilon} = \inf\{k \geq 0,\ M_{n,\varepsilon}^2(k) = 0\}$. Then all the information we need is contained in $(M_{n,\varepsilon}(0), \ldots, M_{n,\varepsilon}(K_{n,\varepsilon}))$: the $K_{n,\varepsilon}$ first values $M_{n,\varepsilon}(0)$ to $M_{n,\varepsilon}(K_{n,\varepsilon} - 1)$, which have second coordinate 1 a.s., describe the law of the successive levels where a mutation occurred on a lineage i up to level $\tau - \varepsilon$. The random variable $M_{n,\varepsilon}^1(K_{n,\varepsilon})$ has the law of the coalescence time between the two consecutive extant individuals $i - 1$ and i at level τ, and $M_{n,\varepsilon}^2(K_{n,\varepsilon}) = 0$ a.s.

Proof of Proposition 3.7
We denote by $\mathbb{T}_{n,n\tau}$ the truncation of $\mathbb{T}_n$ up to level $n\tau$, and by $(Z_{n,n\tau}, Z_{n,n\tau}^{\mathrm{M}})$ the JCCP of $\mathbb{T}_{n,n\tau}$. We define

$$(\tilde{Z}_{n,\tau}(t), \tilde{Z}_{n,\tau}^{\mathrm{M}}(t))_{t\geq 0} := \left(\frac{1}{n} Z_{n,n\tau}(d_n t), Z_{n,n\tau}^{\mathrm{M}}(d_n t)\right)_{t\geq 0},$$

which is in fact, up to a rescaling of time, the JCCP of the rescaled marked splitting tree $\tilde{\mathbb{T}}_n$, truncated up to level τ.
The following lemma is a key tool for the analysis of the genealogy. See Figure 5 for graphical interpretation of some of the objects involved.

Lemma 3.8. *Fix $n \geq 1$ and $\varepsilon > 0$. Define :*

$$\begin{aligned} &t_n^{(0)} := \inf\{t \geq 0,\ \tilde{Z}_{n,\tau}(t) = \tau\}, \quad \text{and for } i \in \mathbb{N},\ t_n^{(i)} := \inf\{t > t_n^{(i-1)},\ \tilde{Z}_{n,\tau}(t) = \tau\}, \\ &S_n^{(i)} := \sup\{t \in [t_n^{(i-1)}, t_n^{(i)}],\ \tilde{Z}_{n,\tau}(t) < \tau - \varepsilon\}, \\ &\text{and } (\Upsilon_n^{(i)}, \Upsilon_n^{(i)\mathrm{M}}) := (\tau - \tilde{Z}_{n,\tau}(S_n^{(i)}-), \Delta\tilde{Z}_{n,\tau}^{\mathrm{M}}(S_n^{(i)})). \end{aligned}$$

Only the first I_n values in the sequence $(t_n^{(i)})_{i\geq 0}$ are finite, and the reversed killed paths

$$e_n^{(i)} := \left\{\left(\tau - \tilde{Z}_{n,\tau}((t_n^{(i)} - t)-), \tilde{Z}_{n,\tau}^{\mathrm{M}}(t_n^{(i)}) - \tilde{Z}_{n,\tau}^{\mathrm{M}}((t_n^{(i)} - t)-)\right),\ 0 \leq t < t_n^{(i)} - t_n^{(i-1)}\right\},\ 1 \leq i < I_n,$$

are i.i.d. Besides, defining for all $1 \leq i < I_n$,

$$e_{n,\varepsilon}^{(i)} := (e_n^{(i)}(t),\ t_n^{(i)} - S_n^{(i)} \leq t < t_n^{(i)} - t_n^{(i-1)}),$$

conditional on $(\Upsilon_n^{(i)}, \Upsilon_n^{(i)\mathrm{M}}) = (x, q)$, $e_{n,\varepsilon}^{(i)}$ has the law of $(\tilde{Z}_n, \tilde{Z}_n^{\mathrm{M}})$, starting at (x, q), conditioned on $\tilde{Z}_n$ hitting 0 before (τ, ∞), and killed when $\tilde{Z}_n$ hits 0.

Proof :
From Theorem 4.3 in [Lam10] which characterizes the law of the JCCP of $\mathbb{T}_{n,\tau}$ (without marks), we deduce that the paths $\{\tilde{Z}_{n,\tau}(t),\ t_n^{(i-1)} \leq t < t_n^{(i)}\}$, $1 \leq i < I_n$, are i.i.d and distributed

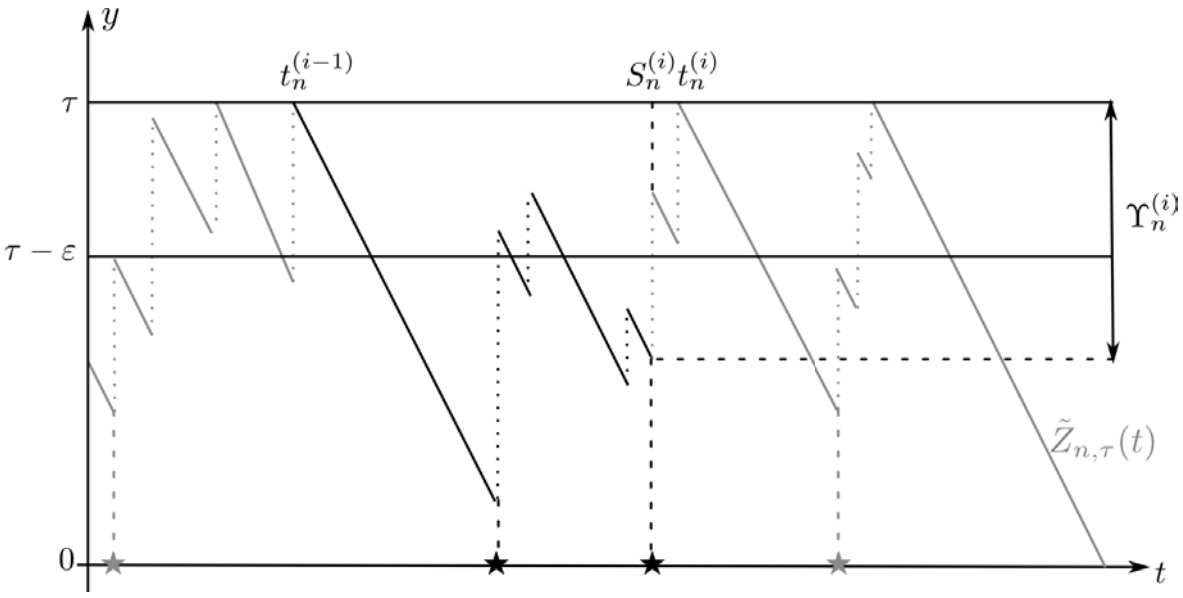

FIGURE 5 – A representation of the (rescaled in time) JCCP $(\tilde{Z}_{n,\tau}, \tilde{Z}^{\mathrm{M}}_{n,\tau})$ (where as before, $\tilde{Z}^{\mathrm{M}}_{n,\tau}$ is represented by the sequence of its jump times, symbolized by stars on the horizontal axis). Here $\Upsilon_n^{(i)\mathrm{M}} = 1$. The reversed path $e^{(i)}_{n,\varepsilon}$ can be read from the black path and black stars, reading the figure upside down and changing y on the vertical axis into $\tau - y$.

as $\tilde{Z}_n$ starting from τ, conditioned on hitting (τ, ∞) before 0 and killed when hitting (τ, ∞). Adapting this property to our marked trees, the i.i.d. property of $\{e_n^{(i)},\ 1 \leq i < I_n\}$ is now straightforward, and the second part of the lemma is then obtained either from an appeal to Proposition 0.7 along with the Markov property of $(H_n^+, H_n^{\mathrm{M}})$ at $L_n(T_n^{(\varepsilon,\infty)}-)$, or using directly a time reversal argument at the last exit time $S_n^{(i)}$ (see [Nag64, Th.3.10]). □

Proof of Proposition 3.7 :
To begin with, we deduce from [Lam10, Corollary 3.5] that for $1 \leq i < I_n$, the set of levels at which birth events occurred on the i-th lineage is the set of the values taken by the future infimum of the rescaled JCCP between $t_n^{(i-1)}$ and $t_n^{(i)}$, i.e. by the process

$$j_n(t) := \inf_{[t, t_n^{(i)}]} \tilde{Z}_{n,\tau}, \quad t_n^{(i-1)} \leq t \leq t_n^{(i)}.$$

As a consequence, the subset of those levels corresponding to 1-type birth events is a.s. equal to $\{j_n(t-),\ t \in J_n^\star\}$, where $J_n^\star$ is the set of jump times of j_n (which are necessarily jump times of $\tilde{Z}_{n,\tau}$) carrying a mark :

$$J_n^\star := \{s \in (t_n^{(i-1)}, t_n^{(i)}],\ \Delta j_n(s) > 0 \text{ and } \Delta \tilde{Z}^{\mathrm{M}}_{n,\tau}(s) > 0\}.$$

Moreover from [Lam10, Theorem 3.4], the coalescence time between lineage i and lineage $i-1$ is given by $\tau - \inf_{[t_n^{(i-1)}, t_n^{(i)}]} \tilde{Z}_{n,\tau} = \tau - j_n(t_n^{(i-1)})$. This yields $\sigma_n^{(i)} = \delta_{(\tau - j_n(t_n^{(i-1)}),0)} + \sum_{t \in J_n^\star} \delta_{(\tau - j_n(t-),1)}$ a.s.

We are interested in the trace on $[\varepsilon, \tau) \times \{0,1\}$ of $\sigma_n^{(i)}$. From the preceding observations and using Lemma 3.8, we deduce the following : first, the point measures $\sigma^{(i)}_{n,\varepsilon}$ are i.i.d. Second, since $p_{n,\varepsilon} = \mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)} \,|\, T_n^{(0,\infty)} < T_n^{-\tau})$, then with probability $1 - p_{n,\varepsilon}$, the infimum of the

excursion $e_n^{(i)}$ is greater than $-\varepsilon$, implying $\sigma_{n,\varepsilon}^{(i)}([\varepsilon,\tau)\times\{0,1\})=0$. Else with probability $p_{n,\varepsilon}$, the point measure $\sigma_{n,\varepsilon}^{(i)}$ has at least one atom.

Conditional on $\sigma_{n,\varepsilon}^{(i)}$ having at least one atom, we choose to order these atoms as in the definition of the space $\mathcal{M}_{\mathrm{P}}$, i.e. increasingly w.r.t. the first coordinate and decreasingly w.r.t. the second one. First note that the reversed future infimum $\big(\tau - j_n((t_n^{(i)}-t)-),\ 0\le t<t_n^{(i)}-t_n^{(i-1)}\big)$ is a.s. equal to the running supremum of $e_n^{(i)}$. Then, from Lemma 3.8 and the first part of this proof, we deduce the following :

- Denote by (a_0,q_0) the first atom of $\sigma_{n,\varepsilon}^{(i)}$. Conditional on $(\Upsilon_n^{(i)},\Upsilon_n^{(i)\mathrm{M}})=(u,q)$, if $q=1$ we have $(a_0,q_0)=(u,1)$ a.s. If $q=0$, then $(a_0,q_0)\in u+\mathrm{d}v\times\{1\}$ with probability $\nu_n^{\mathrm{M}}(u,\mathrm{d}v)$, and $(a_0,q_0)\in u+\mathrm{d}v\times\{0\}$ with probability $\nu_n^{\mathrm{D}}(u,\mathrm{d}v)$. Consequently, (a_0,q_0) is distributed as $M_\varepsilon(0)$.

- Now conditional on (a_0,q_0), if $q_0=0$, then $\sigma_{n,\varepsilon}^{(i)}$ has one unique atom. Now $M_{n,\varepsilon}^2(0)=0$ implies $K_{n,\varepsilon}=0$ a.s., so that we have as announced $\sigma_{n,\varepsilon}^{(i)}\overset{\mathcal{L}}{=}\sum_{k=0}^{K_{n,\varepsilon}}\delta_{M_{n,\varepsilon}(k)}$. Else if $q_0=1$, applying the strong Markov property to (H_n^+,H_n^{M}) at e_n, the next atom of $\sigma_{n,\varepsilon}^{(i)}$ has the law of $M_{n,\varepsilon}(1)$ conditional on $M_{n,\varepsilon}(0)=(x,1)$.

Finally, through a recursive application of the Markov property, stopped the first time an atom has second coordinate 0, we obtain the announced equality in law. □

3.2.2 Limiting Markov chain

Similarly as in the last subsection, for fixed $\varepsilon\in(0,\tau)$ we define a Markov chain M_ε, towards which the sequence $(M_{n,\varepsilon})$ will converge in distribution. First define thanks to Theorem 3.3

$$\mathrm{e}:=\inf\{t\ge 0,\ H^{\mathrm{M}}(t)=1\}.$$

Note that as e_n, e follows an exponential distribution, whose parameter λ is equal to θ in the case of Assumption B.1, and to $\mu(\mathbb{R}_+^*,\{1\})+\rho$ in the case of Assumption B.2.

Then for all $x>0$, $u>0$ and $q\in\{0,1\}$, we set

$$\nu^{\mathrm{M}}(x,\mathrm{d}u):=\mathbb{P}_0(H^+(\mathrm{e})\in\mathrm{d}u, L^{-1}(\mathrm{e})<T^{-x}\,|\,T^{-x}<T^{(\tau-x,\infty)})$$
$$\nu^{\mathrm{D}}(x,\mathrm{d}u):=\mathbb{P}_0(\bar{Z}(T^{-x})\in\mathrm{d}u, L^{-1}(\mathrm{e})\ge T^{-x}\,|\,T^{-x}<T^{(\tau-x,\infty)}).$$

We now want to define $\nu_\varepsilon^{\mathrm{INIT}}$, the counterpart in the limit of the measure $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ defined at rank n. The limiting process Z having infinite variation, this measure will necessarily be described in terms of excursions.
Let $\epsilon\in\mathscr{E}'$ satisfying $-\inf\epsilon\in(\varepsilon,\tau)$. We define

$$S^\varepsilon:=\sup\{t\le\zeta,\ \ \epsilon(t)<-\varepsilon\}$$

the last exit time of ϵ away from $(-\infty,-\varepsilon)$. We then set

$$\Upsilon^\varepsilon(\epsilon):=-\epsilon(S^\varepsilon-),\ \text{ and }\ \Delta\Upsilon^\varepsilon(\epsilon):=\epsilon(S^\varepsilon)-\epsilon(S^\varepsilon-).$$

Recall that the bivariate Lévy process $(\tilde{Z}_n,\tilde{Z}_n^{\mathrm{M}})$ does not converge in general, as observed in Remark 2.1. Then defining a process of marked excursions in the limit is not possible, and for

this reason we do not directly define the counterpart of the r.v. Υ_n^{M}.

In the sequel, when ε is fixed, the notation Υ (resp. $\Delta\Upsilon$) stands for $\Upsilon^\varepsilon(\epsilon)$ (resp. $\Delta\Upsilon^\varepsilon(\epsilon)$). Then if we consider an excursion of Z away from 0 conditioned on hitting level $-\varepsilon$, Υ is the value of Z before its last entry into $(-\varepsilon, \infty)$ (see Figure 4 for a representation in finite variation). Finally, we define for all $(u, q) \in [\varepsilon, \tau) \times \{0, 1\}$:

$$\nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}u, \mathrm{d}q) := \frac{1}{p_\varepsilon} \int_{(u-\varepsilon,\infty)} N'(\Upsilon \in \mathrm{d}u,\ \Delta\Upsilon \in \mathrm{d}v,\ -\inf \epsilon \in [\varepsilon, \tau))\mathbb{B}_{f(v)}(\mathrm{d}q),$$

where $p_\varepsilon := \frac{1}{W(\varepsilon)} - \frac{1}{W(\tau)}$. According to lemma 9 in [OP09], we have $p_\varepsilon = N'(-\inf \epsilon \in (\varepsilon, \tau))$, so that $\nu_\varepsilon^{\mathrm{INIT}}$ is a probability measure.

Next let $M_\varepsilon = (M_\varepsilon(k))_{k\in\mathbb{Z}_+}$ be the Markov chain with values in $[\varepsilon, \tau) \times \{0, 1\}$, defined by :

- For all $k \in \mathbb{Z}_+$, for all $u \geq 0$, conditional on $M_\varepsilon(k) = (x, 1)$,

$$\begin{cases} M_\varepsilon(k+1) \in (x + \mathrm{d}u) \times \{1\} & \text{with probability } \nu^{\mathrm{M}}(x, \mathrm{d}u) \\ M_\varepsilon(k+1) \in (x + \mathrm{d}u) \times \{0\} & \text{with probability } \nu^{\mathrm{D}}(x, \mathrm{d}u) \end{cases}$$

- For all $k \in \mathbb{Z}_+$, conditional on $M_\varepsilon(k) = (x, 0)$, $M_\varepsilon(k+1) = (x, 0)$ a.s.

- For all $u \in [\varepsilon, \tau)$,

$$\begin{cases} \mathbb{P}(M_\varepsilon(0) \in \mathrm{d}u \times \{1\}) = \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}u \times \{1\}) + \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}x \times \{0\})\nu^{\mathrm{M}}(x, \mathrm{d}u - x) \\ \mathbb{P}(M_\varepsilon(0) \in \mathrm{d}u \times \{0\}) = \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}x, \{0\})\nu^{\mathrm{D}}(x, \mathrm{d}u - x) \end{cases}$$

The values 0 and 1 stand as earlier for the absence or presence of a mutation.
Let K_ε be defined as follows :

$$K_\varepsilon := \inf\{k \geq 0,\ M_\varepsilon^2(k) = 0\}.$$

Under $\mathbb{P}_x(\,\cdot\,|\,T^0 < T^{(\tau,\infty)})$, the interval $[0, L(T^0))$ is a.s. finite, and $K_\varepsilon + 1$ is a.s. equal to the number of jumps of the counting process H^{M} on this interval, so that K_ε is a.s. finite.

The main argument needed for the proof of Theorems 2.2 and 2.7 is given by the following proposition :

Proposition 3.9. *For all $k \geq 0$, as $n \to \infty$, the $(k+1)$-tuple $(M_{n,\varepsilon}(0), ...M_{n,\varepsilon}(k))$ converges in distribution towards $(M_\varepsilon(0), ...M_\varepsilon(k))$.*

For now we admit this proposition and relegate its proof to Section 3.3. We now have all the necessary ingredients to prove our main theorem.

3.2.3 Proof of Theorems 2.2 and 2.7

In this Section we assume that one of the two Assumptions B.1 or B.2 is satisfied. We first establish the convergence of Σ_n towards a Poisson point measure with intensity $\mathrm{Leb} \otimes \Pi$, making use of the law of rare events for null arrays (see e.g. Theorem 16.18 in [Kal02]). The proof of Theorem 2.7, which is valid both under B.1 and B.2, will then consist in identifying the intensity measures Π_2 with the measure Π.

Our main objects of interest in this section are then the point measures $\Sigma_n = \sum_{i=1}^{I_n} \delta_{(\frac{in}{d_n}, \sigma_n^{(i)})}$, where we recall that the random variables $\sigma_n^{(i)}$ have values in the space $\mathcal{M}_{\mathrm{P}}$ defined in Section 2.2.3. The trace σ-field on $\mathcal{M}_{\mathrm{P}}$ is in particular generated by the class $\big\{p_B,\ B = [\varepsilon,\tau)\times\{0\},\ B = [\varepsilon,\tau)\times\{1\}\big\}_{\varepsilon\in(0,\tau)}$. Note that a measure $\delta_{(a_0,0)} + \sum_{i=1}^{j} \delta_{(a_i,1)}$ in $\mathcal{M}_{\mathrm{P}}$ is characterized by the set of first coordinates of its atoms $\{a_0,\dots,a_j\}$. Then, if we denote by $B_{m,\varepsilon}$ the subset of $\mathcal{M}_{\mathrm{P}}$ defined by

$$B_{m,\varepsilon} = \{\sigma \in \mathcal{M}_{\mathrm{P}},\ \sigma([\varepsilon,\tau)\times\{0,1\}) = m+1\},$$

the class $\mathscr{C} := \{B_{m,\varepsilon},\ m\in\mathbb{Z}_+,\ \varepsilon\in(0,\tau)\}$ is a generating class for the trace σ-field on $\mathcal{M}_{\mathrm{P}}$.

Proposition 3.10. *The sequence (Σ_n) converges in distribution towards a Poisson point measure Σ on $[0,1]\times\mathcal{M}_{\mathrm{P}}$ with intensity measure $\mathrm{Leb}\otimes\Pi$, where Π is a measure on $\mathcal{M}_{\mathrm{P}}$ characterized as follows : for all $m\in\mathbb{Z}_+$ and $\varepsilon\in(0,\tau)$,*

$$\Pi(B_{m,\varepsilon}) = p_\varepsilon \mathbb{P}(K_\varepsilon = m).$$

Proof of Proposition 3.10 :
To begin with, we prove that as $n\to\infty$, $\mathbb{E}(\Sigma_n(B\times C)) \to \mathbb{E}(\Sigma(B\times C))$ for any Borel set B in $[0,1]$ and any measurable set C of $\mathcal{M}_{\mathrm{P}}$. From Lemma 3.8, we know that the point measures $\sigma_n^{(i)}$, $1\le i < I_n$, are independent, yielding

$$\begin{aligned}\mathbb{E}\left(\Sigma_n(B\times C)\right) &= \sum_{i=1}^{I_n-1} \mathbb{P}\left(\frac{in}{d_n}\in B,\ \sigma_n^{(i)}\in C\right)\\ &= \left(\frac{d_n}{n}\mathbb{P}(\sigma_n^{(1)}\in C)\right)\left(\frac{n}{d_n}\sum_{i=1}^{I_n-1}\mathbb{1}_{\frac{in}{d_n}\in B}\right)\end{aligned}$$

Recall that we assumed that $I_n \underset{n\to\infty}{\sim} \frac{d_n}{n}$. The second term in the right-hand side clearly converges in distribution towards $\mathrm{Leb}(B)$, and it remains to prove the convergence of the first term. Now using a monotone class argument, it suffices to prove this convergence for sets C in the class $\mathscr{C}$ defined above.

For all $\varepsilon\in(0,\tau)$ and $m\in\mathbb{Z}_+$, we have by definition of $\sigma_{n,\varepsilon}^{(1)}$ and according to Proposition 3.7 :

$$\frac{d_n}{n}\mathbb{P}(\sigma_n^{(1)}\in B_{m,\varepsilon}) = \frac{d_n}{n}\mathbb{P}(\sigma_{n,\varepsilon}^{(1)}\in B_{m,\varepsilon}) = \frac{d_n}{n}p_{n,\varepsilon}\mathbb{P}(K_{n,\varepsilon}=m).$$

First for $m=0$, we then have

$$\begin{aligned}\frac{d_n}{n}\mathbb{P}(\sigma_n^{(1)}\in B_{0,\varepsilon}) &= \frac{d_n}{n}p_{n,\varepsilon}\mathbb{P}(M_{n,\varepsilon}^2(0)=0)\\ &\underset{n\to\infty}{\longrightarrow} p_\varepsilon\mathbb{P}(M_\varepsilon^2(0)=0) \ = \ p_\varepsilon\mathbb{P}(K_\varepsilon=0),\end{aligned}$$

and for $m\ge 1$,

$$\begin{aligned}\frac{d_n}{n}\mathbb{P}(\sigma_n^{(1)}\in B_{m,\varepsilon}) &= \frac{d_n}{n}p_{n,\varepsilon}\mathbb{P}(M_{n,\varepsilon}^2(m-1)=1,\ M_{n,\varepsilon}^2(m)=0)\\ &\underset{n\to\infty}{\longrightarrow} p_\varepsilon\mathbb{P}(M_\varepsilon^2(m-1)=1,\ M_\varepsilon^2(m)=0) \ = \ p_\varepsilon\mathbb{P}(K_\varepsilon=m),\end{aligned}$$

where the convergences are obtained from an appeal to Proposition 3.9 and using the fact that $\frac{d_n}{n}p_{n,\varepsilon}\to p_\varepsilon$.

Finally, we get for all $B, C \in \mathcal{B}([0,1]) \times \mathscr{C}$:

$$\mathbb{E}\left(\Sigma_n(B \times C)\right) \underset{n\to\infty}{\to} \Pi(B \times C).$$

The point measures (Σ_n) form a null array of simple point measures on $[0,1] \times \mathcal{M}_{\mathrm{P}}$, therefore, from the conclusion above, the theorem is a straightforward consequence of Corollary 0.3. □

The following lemma is the last step preceding the proof of Theorems 2.2 and 2.7. For $i \geq 1$, we define the sequence $(\mathrm{e}_i)_{i\geq 0}$ as follows : first set $\mathrm{e}_0 = 0$. Then, for $i \geq 1$, e_i denotes the i-th jump time of H^{M} if it exists, or is else set to $+\infty$. Note that e_1 is in fact equal to e a.s. We then define $J := \sup\{i \geq 0,\ \mathrm{e}_i < L(T^0)\}$, which is in particular finite a.s. on $L(T^0) < \infty$.

Lemma 3.11. *For all $m \in \mathbb{Z}_+$ and $\varepsilon \in (0,\tau)$ we have*

$$\mathbb{P}(K_\varepsilon = m) = \int_{[\varepsilon,\tau)\times\{0,1\}} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}x, \mathrm{d}q)\mathbb{P}_x(J = m - q \,|\, T^0 < T^{(\tau,\infty)}). \tag{6}$$

Remark 3.12. *Let σ be defined as in Theorem 2.7. Then for $x, \varepsilon \in (0,\tau)$, $m \in \mathbb{Z}_+$, if $x \geq \varepsilon$ then $\mathbb{P}_x(J = m \,|\, T^0 < T^{(\tau,\infty)})$ is in fact equal to $\mathbb{P}_x(\sigma \in B_{m,\varepsilon} \,|\, T^0 < T^{(\tau,\infty)})$.*

Proof :
Fix $\varepsilon \in (0,\tau)$. First note that for all $x \in [\varepsilon,\tau)$,

$$\mathbb{P}_x(J = 0 \,|\, T^0 < T^{(\tau,\infty)}) = \mathbb{P}_x(\mathrm{e}_1 \geq L(T^0) \,|\, T^0 < T^{(\tau,\infty)}) = \nu^{\mathrm{D}}(x, [\varepsilon,\tau)), \tag{7}$$

and

$$\begin{aligned}
&\mathbb{P}_x(J = 1 \,|\, T^0 < T^{(\tau,\infty)}) \\
&= \mathbb{P}_x(\mathrm{e}_1 < L(T^0),\ \mathrm{e}_2 \geq L(T^0) \,|\, T^0 < T^{(\tau,\infty)}) \\
&= \int_{[0,\tau-x)} \mathbb{P}_x(\mathrm{e}_1 < L(T^0),\ \mathrm{e}_2 \geq L(T^0),\ H^+(\mathrm{e}_1) \in x + \mathrm{d}u,\ T^0 < T^{(\tau,\infty)})/\mathbb{P}_x(T^0 < T^{(\tau,\infty)}) \\
&= \int_{0,\tau-x)} \mathbb{P}_x(\mathrm{e}_1 < L(T^0),\ H^+(\mathrm{e}_1) \in x + \mathrm{d}u)\mathbb{P}_{x+u}(\mathrm{e}_1 \geq L(T^0),\ T^0 < T^{(\tau,\infty)})/\mathbb{P}_x(T^0 < T^{(\tau,\infty)}) \\
&= \int_{[0,\tau-x)} \frac{\mathbb{P}_x(\mathrm{e}_1 < L(T^0),\ H^+(\mathrm{e}_1) \in x + \mathrm{d}u,\ T^0 < T^{(\tau,\infty)})}{\mathbb{P}_x(T^0 < T^{(\tau,\infty)})} \frac{\mathbb{P}_{x+u}(\mathrm{e}_1 \geq L(T^0),\ T^0 < T^{(\tau,\infty)})}{\mathbb{P}_{x+u}(T^0 < T^{(\tau,\infty)})} \\
&= \int_{[0,\tau-x)} \mathbb{P}_x(\mathrm{e}_1 < L(T^0),\ H^+(\mathrm{e}_1) \in x + \mathrm{d}u \,|\, T^0 < T^{(\tau,\infty)})\mathbb{P}_{x+u}(\mathrm{e}_1 \geq L(T^0) \,|\, T^0 < T^{(\tau,\infty)}),
\end{aligned}$$

where in the third equality we applied the Markov property to (H^+, H^{M}) at the stopping time e_1. We omit for now to justify properly this application of the Markov property : details on filtrations and stopping times are provided in Section 3.3.1 (see Proposition 3.20). Finally, this gives :

$$\mathbb{P}_x(J = 1 \,|\, T^0 < T^{(\tau,\infty)}) = \int_{u\in[0,\tau-x)} \nu^{\mathrm{M}}(x, \mathrm{d}u)\nu^{\mathrm{D}}(x + u, [\varepsilon,\tau)). \tag{8}$$

We first show (6) for $m = 0$. Since $J \geq 0$ a.s., from (7) we have

$$\begin{aligned}
\int_{[\varepsilon,\tau)\times\{0,1\}} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}x, \mathrm{d}q)\mathbb{P}_x(J = -q \,|\, T^0 < T^{(\tau,\infty)}) &= \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}x, \{0\})\nu^{\mathrm{D}}(x, [\varepsilon,\tau)) \\
&= \mathbb{P}(M_\varepsilon^2(0) = 0) = \mathbb{P}(K_\varepsilon = 0).
\end{aligned}$$

Similarly we prove (6) for $m = 1$, using (7) and (8) in the second equality :

$$\begin{aligned}
&\int_{[\varepsilon,\tau)\times\{0,1\}} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}x, \mathrm{d}q)\mathbb{P}_x(J = 1 - q \,|\, T^0 < T^{(\tau,\infty)}) \\
&= \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}x, \{0\})\mathbb{P}_x(J = 1 \,|\, T^0 < T^{(\tau,\infty)}) + \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}v, \{1\})\mathbb{P}_v(J = 0 \,|\, T^0 < T^{(\tau,\infty)}) \\
&= \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}x, \{0\}) \left(\int_{u\in[0,\tau-x)} \nu^{\text{M}}(x, \mathrm{d}u)\nu^{\text{D}}(x + u, [\varepsilon, \tau)) \right) + \int_{[\varepsilon,\tau)} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}v, \{1\})\nu^{\text{D}}(v, [\varepsilon, \tau)) \\
&= \int_{v\in[\varepsilon,\tau)} \left(\nu_\varepsilon^{\text{INIT}}(\mathrm{d}v, \{1\}) + \int_{x\in[\varepsilon,v)} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}x, \{0\})\nu^{\text{M}}(x, \mathrm{d}v - x) \right) \nu^{\text{D}}(v, [\varepsilon, \tau)) \\
&= \int_{[\varepsilon,\tau)\times\{0,1\}} \mathbb{P}(M_\varepsilon(0) \in \mathrm{d}u \times \{1\})\mathbb{P}(M_\varepsilon^2(1) = 0 \,|\, M_\varepsilon(0) = (u, 1)) \\
&= \mathbb{P}(K_\varepsilon = 1).
\end{aligned}$$

It is then clear by induction on m that (6) is true for all $m \in \mathbb{Z}_+$, which ends the proof. □

In the proof below, we use Proposition 3.10 and Lemma 3.11 to deduce Theorem 2.7, which is in fact also valid both under B.1 and B.2. Theorem 2.2 is then simply a consequence of Theorem 2.7, using the independence between H^+ and H^{M} that arises under Assumption B.1.

Proof of Theorem 2.7 :
Fix $m \in \mathbb{Z}_+$ and $\varepsilon \in (0, \tau)$. First, from Proposition 3.10, along with Lemma 3.11 and Remark 3.12, we deduce

$$\Pi(B_{m,\varepsilon}) = p_\varepsilon \int_{[\varepsilon,\tau)\times\{0,1\}} \nu_\varepsilon^{\text{INIT}}(\mathrm{d}x, \mathrm{d}q)\mathbb{P}_x(\sigma \in B_{m-q,\varepsilon} \,|\, T^0 < T^{(\tau,\infty)}). \tag{9}$$

We now want to prove that Π and Π_2 coincide on the generating class $\mathscr{C}$, using (9). We denote by σ the point measure $\Psi(H^+, \epsilon^{\text{M}} + H^{\text{M}}, L(T^0))$ that appears in the statement of the theorem, and we consider

$$\Pi_2(B_{m,\varepsilon}) = N''(\sigma \in B_{m,\varepsilon},\ \sup \epsilon < \tau).$$

Recall first that any point measure belonging to $B_{m,\varepsilon}$ necessarily has at least one atom with first coordinate greater than ε. Using the (slightly abusive) notation $T^{(\varepsilon,\infty)}$ for the first entrance time in (ε, ∞) of an excursion $\epsilon \in \mathscr{E}''$, we apply the Markov property to (H^+, H^{M}) at $L(T^{(\varepsilon,\infty)})$: recall that $H^+(L(T^{(\varepsilon,\infty)})) = Z(T^{(\varepsilon,\infty)})$, and that σ might have an atom coming from a jump of H^{M} at $L(T^{(\varepsilon,\infty)})$. Conditional on $\Delta\epsilon(T^{(\varepsilon,\infty)}) = v$, this occurs with probability $f(v)$. Again, see

Section 3.3.1 for details about filtrations and stopping times. This gives :

$$
\begin{aligned}
&N''(\sigma \in B_{m,\varepsilon},\ \sup\epsilon < \tau)\\
&= \int_{[\varepsilon,\tau)} \int_{[u-\varepsilon,\infty)} N''(\sigma \in B_{m,\varepsilon},\ \sup\epsilon \in [\varepsilon,\tau),\ \epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}u,\ \Delta\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}v)\\
&= \int_{[\varepsilon,\tau)\times\{0,1\}} \int_{[u-\varepsilon,\infty)} N''(\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}u,\ \Delta\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}v)\ \mathbb{B}_{f(v)}(\mathrm{d}q)\\
&\qquad\qquad \times \mathbb{P}_u(\sigma \in B_{m-q,\varepsilon},\ T^0 < T^{(\tau,\infty)})\\
&= \int_{[\varepsilon,\tau)\times\{0,1\}} \int_{[u-\varepsilon,\infty)} N''(\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}u,\ \Delta\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}v)/\mathbb{P}_u(T^0 < T^{(\tau,\infty)})\ \mathbb{B}_{f(v)}(\mathrm{d}q)\\
&\qquad\qquad \times \mathbb{P}_u(\sigma \in B_{m-q,\varepsilon},\ T^0 < T^{(\tau,\infty)})\mathbb{P}_u(T^0 < T^{(\tau,\infty)})\\
&= \int_{[\varepsilon,\tau)\times\{0,1\}} \int_{[u-\varepsilon,\infty)} N''(\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}u,\ \Delta\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}v,\ \sup\epsilon \in [\varepsilon,\tau))\ \mathbb{B}_{f(v)}(\mathrm{d}q)\\
&\qquad\qquad \times \mathbb{P}_u(\sigma \in B_{m-q,\varepsilon}\,|\,T^0 < T^{(\tau,\infty)}).
\end{aligned}
$$

Now from the definition of N'', we know that

$$N''(\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}u,\ \Delta\epsilon(T^{(\varepsilon,\infty)}) \in \mathrm{d}v,\ \sup\epsilon \in [\varepsilon,\tau)) = N'(\Upsilon \in \mathrm{d}u,\ \Delta\Upsilon \in \mathrm{d}v,\ -\inf\epsilon \in [\varepsilon,\tau)),$$

which entails

$$N''(\sigma \in B_{m,\varepsilon},\ \sup\epsilon < \tau) = p_\varepsilon \int_{[\varepsilon,\tau)\times\{0,1\}} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}u,\mathrm{d}q)\mathbb{P}_u(\sigma \in B_{m-q,\varepsilon}\,|\,T^0 < T^{(\tau,\infty)}).$$

This equality, along with (9), leads to the expected result. □

Proof of Theorem 2.2 :
As announced, the latter proof is also valid under B.1, in which case H^{M} is independent from H^+, and is a Poisson process with parameter θ. Moreover, $f \equiv 0$ implies $N''(\epsilon^{\mathrm{M}} = 1) = 0$. Thus Theorem 2.2 can be directly deduced from Theorem 2.7. □

3.2.4 Proof of Proposition 2.6

Finally, we prove here Proposition 2.6. The counterpart of this proposition under B.2 (stated in the second paragraph of Section 2.3) can be established by an easy adaptation of the upcoming proof.

Proof of Proposition 2.6 :
Fix $x \in (0,\tau)$. Consider the process H^+ under $\mathbb{P}_x(\cdot \cap \{T^0 < T^{(\tau,\infty)}\})$, killed at $L(T^0)$. This process is an inhomogeneous killed subordinator, with jump measure denoted by ν^{K}. Hereafter we prove that ν^{K} and μ^{K} coincide.

Let F be a nonnegative continuous $\mathcal{F}_{L^{-1}}$-measurable function on $\mathbb{R}_+ \times (\mathbb{R}_+ \cup \{+\infty\})$, and U a $\mathcal{F}_{L^{-1}}$-predictable process. Recalling that $L(T^0)$ is a $\mathcal{F}_{L^{-1}}$-stopping time, we have by compensation formula for any fixed $t > 0$:

$$
\begin{aligned}
&\mathbb{E}_x\left(\sum_{0<r\le t\wedge L(T^0)} \left(\mathbb{1}_{\Delta H_r^+>0}\, U_r\, F(H_{r-}^+, \Delta H_r^+)\right),\ T^0 < T^{(\tau,\infty)}\right)\\
&\qquad = \mathbb{E}_x\left(\int_0^{t\wedge L(T^0)} \mathrm{d}s U_s \int_{(0,+\infty]} F(H_s^+, z)\nu^{\mathrm{K}}(H_s^+, \mathrm{d}z)\right) \quad (10)
\end{aligned}
$$

where $\Delta H_r^+ := +\infty$ if $H_r^+ = +\infty$, and $\Delta H_r^+ := H_r^+ - H_{r-}^+$ otherwise.

On the other hand, we have :

$$\begin{aligned}
&\mathbb{E}_x\left(\sum_{0<r\le t\wedge L(T^0)}\left(\mathbb{1}_{\Delta H_r^+>0}\,U_r\,F(H_{r-}^+,\Delta H_r^+)\right),\ T^0<T^{(\tau,\infty)}\right)\\
&=\mathbb{E}_x\Bigg(\sum_{0<r\le t}\mathbb{E}\Big(U_r\,F(H_{r-}^+,\Delta H_r^+),\ \Delta H_r^+>0,\ r<L(T^0),\ T^0<T^{(\tau,\infty)}\,|\,\mathcal{F}_{L^{-1}(r)}\Big)\\
&\qquad\qquad+U_r\,F(H_{r-}^+,\Delta H_r^+),\ \Delta H_r^+>0,\ r=L(T^0),\ T^0<T^{(\tau,\infty)}\Bigg)\\
&=\mathbb{E}_x\left(\sum_{0<r\le t\wedge L(T^0)}U_r\,F(H_{r-}^+,\Delta H_r^+)\,\left(\mathbb{1}_{H_{r-}^+<H_r^+<\tau}\,\mathbb{P}_{H_r^+}(T^0<T^{(\tau,\infty)})+\mathbb{1}_{H_{r-}^+<\tau,\ \Delta H_r^+=+\infty}\right)\right),
\end{aligned}$$

using on the one hand the $\mathcal{F}_{L^{-1}(r)}$-measurability of every term but $\mathbb{1}_{T^0<T^{(\tau,\infty)}}$ in the conditional expectation and the Markov property at time r, and on the other hand the fact that $\{r = L(T^0),\ T^0<T^{(\tau,\infty)}\}$ and $\{H_{r-}^+<\tau,\ \Delta H_r^+=+\infty\}$ coincide under $\mathbb{E}_x$.
We now express the sum in the right hand side in terms of excursions.

$$\begin{aligned}
&\mathbb{E}_x\left(\sum_{0<r\le t\wedge L(T^0)}\left(\mathbb{1}_{\Delta H_r^+>0}\,U_r\,F(H_{r-}^+,\Delta H_r^+)\right),\ T^0<T^{(\tau,\infty)}\right)\\
&=\mathbb{E}_x\Bigg(\sum_{0\le g<L^{-1}(t)\wedge T^0}U_{L(g)}\,F(H_{L(g)-}^+,e_g(\zeta))\mathbb{1}_{\{-\inf e_g<H_{L(g)-}^+<\tau-e_g(\zeta)\}}\,\mathbb{P}_{H_{L(g)}^+}(T^0<T^{(\tau,\infty)})\\
&\qquad\qquad+U_{L(g)}\,F(H_{L(g)-}^+,+\infty)\,\mathbb{1}_{\{-\inf e_g\ge H_{L(g)}^+\}}\Bigg),
\end{aligned}$$

where the sum in the right-hand side is taken over all the left-end points of excursions intervals. Then by compensation formula,

$$\begin{aligned}
&\mathbb{E}_x\left(\sum_{0<r\le t\wedge L(T^0)}\left(\mathbb{1}_{\Delta H_r^+>0}\,U_r\,F(H_{r-}^+,\Delta H_r^+)\right),\ T^0<T^{(\tau,\infty)}\right)\\
&=\mathbb{E}_x\Bigg(\int_0^{t\wedge L(T^0)}\mathrm{d}s\,U_s\Bigg(\int_{(0,\tau-H_s^+)}F(H_s^+,z)\,\mathbb{P}_{H_s^++z}(T^0<T^{(\tau,\infty)})N(\epsilon(\zeta)\in\mathrm{d}z,\,-\inf\epsilon<H_s^+)\\
&\qquad\qquad+\,F(H_s^+,+\infty)N(-\inf\epsilon\ge H_s^+)\Bigg)\Bigg)\qquad(11)
\end{aligned}$$

Finally from (10) and (11) we deduce that for all $a\in(0,\tau)$, $z\in(0,\infty]$,

$$\nu^{\mathrm{K}}(a,\mathrm{d}z)=\mathbb{1}_{z<\tau-a}\mathbb{P}_{a+z}(T^0<T^{(\tau,\infty)})N(\epsilon(\zeta)\in\mathrm{d}z,\,-\inf\epsilon<a)+N(-\inf\epsilon\ge a)\delta_{+\infty}(\mathrm{d}z),$$

which yields, using Proposition 0.7 and the fact that $N(-\inf\epsilon\ge a)=\frac{1}{W(a)}$,

$$\nu^{\mathrm{K}}(a,\mathrm{d}z)=\frac{W(\tau-a-z)}{W(\tau)}\int_0^a\mathrm{d}x\frac{W(a-x)}{W(a)}\Lambda(x+\mathrm{d}z)+\frac{1}{W(a)}\delta_{+\infty}(\mathrm{d}z)=\mu^{\mathrm{K}}(a,\mathrm{d}z).$$

From this result we deduce that under $\mathbb{P}_x$, H^{K} has the law of H^+ under $\mathbb{P}_x(\,\cdot\cap\{T^0<T^{(\tau,\infty)}\})$, killed at $L(T^0)$, which finishes the proof. □

3.3 Convergence of the Markov chains

3.3.1 Weak convergence of ν_n^{M} towards ν^{M} and characterization of these measures

Before proving the convergence in law of $M_{n,\varepsilon}$ to M_ε, we show in this subsection that the sequence of measure (ν_n^{M}) converges weakly towards ν^{M}. Recall that $\nu_n^{\mathrm{M}}(x,\cdot)$ is the law of the amount of time elapsed between two mutations conditional on the latest one to have happened at level $\tau - x$:

$$\nu_n^{\mathrm{M}}(x,\mathrm{d}u) := \mathbb{P}_0\big(H_n^+(\mathrm{e}_n) \in \mathrm{d}u, L_n^{-1}(\mathrm{e}_n) < T_n^{-x} \,|\, T_n^{-x} < T_n^{(\tau-x,\infty)}\big).$$

The announced weak convergence of ν_n^{M} towards ν^{M} is contained in the following result, which also gives an expression of these measures. Recall that in case Z drifts to $-\infty$, we denoted by $k = \frac{1}{W(\infty)}$ the killing rate of (H^+, H^{M}). If Z does not drift to $-\infty$, we set $k = 0$.

Theorem 3.13. *For all z, y in $\mathbb{R}_+$ such that $z + y \leq \tau - x$, the measure*

$$\mathbb{P}(H_n^+(\mathrm{e}_n-) \in \mathrm{d}z, \Delta H_n^+(\mathrm{e}_n) \in \mathrm{d}y, L_n^{-1}(\mathrm{e}_n) < T_n^{-x} \,|\, T_n^{-x} < T_n^{(\tau-x,\infty)})$$

converges weakly towards

$$\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z, \Delta H^+(\mathrm{e}) \in \mathrm{d}y, L^{-1}(\mathrm{e}) < T^{-x} \,|\, T^{-x} < T^{(\tau-x,\infty)}).$$

Besides, we have

$$\begin{aligned}
&\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L^{-1}(\mathrm{e}) < T^{-x} < T^{(\tau-x,\infty)}) \\
&\quad = \Bigg\{ \tilde{\mu}(\mathrm{d}y, \{1\}) \Bigg[U_*^{(\lambda+k)}(\mathrm{d}z) - \int_{[0,z)} \pi(\mathrm{d}a) \int_{[a,z)} U_*^{(\lambda+k)}(\mathrm{d}z - b) g^x(a, \{0\}, \mathrm{d}b - a) \Bigg] \\
&\qquad\qquad - \pi(\mathrm{d}z) g^x(z, \{1\}, \mathrm{d}y) \Bigg\} \frac{W(\tau - x - z - y)}{W(x)}, \quad (12)
\end{aligned}$$

where

- *$\tilde{\mu}$ is the Lévy measure of (H^+, H^{M}), yielding $\tilde{\mu}(\mathrm{d}y, \{1\}) = \theta\delta_0(\mathrm{d}y)$ under B.1, and $\tilde{\mu}(\mathrm{d}y, \{1\}) = \mu(\mathrm{d}y, \{1\}) + \rho\delta_0(\mathrm{d}y)$ under B.2.*
- *$U_*^{(l)}$ is the l-resolvent measure of the subordinator H^* defined in Theorem 3.4, that is*

$$U_*^{(l)}(\mathrm{d}z) := \int_{(0,\infty)} e^{-lt} \mathbb{P}(H^*(t) \in \mathrm{d}z)\mathrm{d}t,$$

- *π is a finite measure defined by*

$$\pi(\mathrm{d}z) := \mathbb{P}(H^+(L(T^{-x})-) \in \mathrm{d}z, L(T^{-x}) \leq \mathrm{e})$$

- *and finally,*

$$\begin{aligned}
g^x(a, \mathrm{d}q, \mathrm{d}v) = &\frac{b^2}{2}(W'(x+a) - \eta W(x+a))\delta_0(\mathrm{d}v)\delta_0(\mathrm{d}q) \\
&+ \int_{(0,\infty)} (e^{-\eta u} W(x+a) - W(x+a-u))\, \mathbb{B}_{f(u+v)}(\mathrm{d}q)\, \Lambda(u + \mathrm{d}v)\mathrm{d}u. \quad (13)
\end{aligned}$$

Recall that λ is the parameter of the exponential variable e, and is equal to θ (resp. $\mu(\mathbb{R}_+^*, \{1\})+\rho$) under Assumption B.1 (resp. B.2).

Remark 3.14. *In the case of Assumption B.1 several simplifications can be made : we know that H^+ and H^{M} are independent, and $f \equiv 0$. Then the processes H^+ and H^* are equal in law in $\mathbb{D}(\mathbb{R}_+)$, and further $U_*^{(\lambda)}(\,\cdot\,) = \mathbb{P}(H^+(\mathrm{e}) \in \,\cdot\,)$. Second, we have $g^x(a, \{1\}, \mathrm{d}v) = 0$ for all $x > 0$ and $a, v \geq 0$, and from [Kyp06, (8.29)], we see that $g^x(a, \{0\}, \mathrm{d}v) = \mathbb{P}_{-(x+a)}(Z(T^{(0,\infty)}) \in \mathrm{d}v)$. Finally,* (12) *yields*

$$\begin{aligned}&\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L^{-1}(\mathrm{e}) < T^{-x} < T^{(\tau-x,\infty)})\\ &\quad = \theta\delta_0(\mathrm{d}y)\left[U_*^{(\lambda)}(\mathrm{d}z) - \int_{[0,z)} \pi(\mathrm{d}a) \int_{[a,z)} U_*^{(\lambda)}(\mathrm{d}z-b) g^x(a, \{0\}, \mathrm{d}b-a)\right] \frac{W(\tau-x-z-y)}{W(\tau)},\end{aligned} \tag{14}$$

which will be proven along with the theorem.

Remark 3.15. *The measure π is not explicit, and under Assumption B.2 the random variable* e *is not independent of H^+ and $L(T^{-x})$. However we can give another interpretation of π in terms of a Poisson point measure : define similarly as in Section 3.1.2, for all $t \geq 0$,*

$$\xi(t) := \begin{cases} (e_t(\zeta), -\inf_{(0,\zeta)} e_t, \Delta H^{\text{M}}(t))_{t\geq 0} & \text{if } L^{-1}(t-) < L^{-1}(t) \\ \partial & \text{else} \end{cases},$$

where ∂ is an additional isolated point, and $(t, e_t)_{t\geq 0}$ the excursion process of Z (excursions from the past supremum). Then $(t, \xi(t))_{t\geq 0}$ is a Poisson point process with values in $\mathbb{R}_+ \times (\mathbb{R}_+^*)^2 \times \{0,1\}$. Denote by m its intensity measure, and by ξ^i the i-th coordinate of ξ. Recall that H^+ has drift $\frac{b^2}{2}$ and jump process $(\xi^1(t))$, and define $F(t) := \frac{b^2}{2}t + \sum_{s<t} \xi^1(s)$ for all $t \geq 0$. Then

$$\pi(\mathrm{d}z) = m\left(F(T) \in \mathrm{d}z,\ \{\xi^3(s) = 0\ \forall s < T\}\right),$$

where $T := \inf\{t \geq 0,\ \xi^2(t) > x + F(t)\}$.

We turn our attention to the proof of Theorem 3.13, which will mainly rely on the following proposition.

Proposition 3.16. *The Laplace transform $\mathbb{E}(e^{-rH_n^+(\mathrm{e}_n-)},\ L(T_n^{-x}) < \mathrm{e}_n)$ converges to*

$$\mathbb{E}(e^{-rH^+(\mathrm{e}-)},\ L(T^{-x}) < \mathrm{e}) = \frac{\lambda}{\lambda + k + \psi^*(r)} \int_{[0,\infty)} e^{-ar} \gamma^x(a, 0, r) \pi(\mathrm{d}a),$$

where ψ^* is the Laplace exponent of H^* defined in Theorem 3.4, π and g^x are defined in Theorem 3.13 above, and $\gamma^x(a, q, r) := \int_{[0,\infty)} e^{-rv} g^x(a, \{q\}, \mathrm{d}v)$.

To prove the theorem and proposition above, we will need the following lemmas.

Lemma 3.17. *Define for $a, h, t \in \mathbb{R}_+$:*

$$\pi_n(\mathrm{d}a) := \mathbb{P}(\bar{\bar{Z}}_n(T_n^{-x}) \in \mathrm{d}a,\ L_n(T_n^{-x}) \leq \mathrm{e}_n).$$

Then (π_n) converges weakly towards the measure π defined in Theorem 3.13.

Proof :
To prove the lemma we prove that $(\bar{\tilde{Z}}_n(T_n^{-x}), L_n(T_n^{-x}), \mathrm{e}_n)$ converges in distribution towards $(Z(T^{-x}), L(T^{-x}), \mathrm{e})$. From Theorem 3.6, we know that the triplet $(\tilde{Z}_n, L_n, H_n^{\mathrm{M}})$ converges in distribution towards (Z, L, H^{M}). Using the Skorokhod representation theorem, there exists a sequence $(\tilde{\mathcal{Z}}_n, \mathcal{L}_n, \mathcal{H}_n^{\mathrm{M}})$ converging almost surely towards (Z, L, H^{M}), and such that $(\tilde{\mathcal{Z}}_n, \mathcal{L}_n, \mathcal{H}_n^{\mathrm{M}})$ and $(\tilde{Z}_n, L_n, H_n^{\mathrm{M}})$ are equal in law. We will use the notation $\mathcal{T}_n^{-x}$ for the first entrance time of $\tilde{\mathcal{Z}}_n$ in $\{-x\}$, and $\bar{\mathcal{Z}}_n(t) = \sup_{[0,t]} \tilde{\mathcal{Z}}_n$.

Thanks to Proposition 3.1.(i), we know that as $n \to \infty$, $\mathcal{T}_n^{-x} \to T^{-x}$ a.s. Then note that $\mathcal{T}_n^{-x}$ is a continuity time for $\bar{\mathcal{Z}}_n$ and $\mathcal{L}_n$, since they are a.s. constant in a neighborhood of $\mathcal{T}_n^{-x}$, and hence we get from Proposition 2.1 (b.5) in [JS87] that $\tilde{\mathcal{Z}}_n(\mathcal{T}_n^{-x}) \to Z(T^{-x})$ and $\mathcal{L}_n(\mathcal{T}_n^{-x}) \to L(T^{-x})$ a.s.

We have $\mathcal{E}_n = T^1(\mathcal{H}_n^{\mathrm{M}})$ and $\mathrm{e} = T^1(H^{\mathrm{M}})$, where $\mathcal{H}_n^{\mathrm{M}}$ and H^{M} are Poisson processes satisfying $\mathcal{H}_n^{\mathrm{M}} \xrightarrow[\mathbb{P}]{a.s.} H^{\mathrm{M}}$. Here Proposition VI.2.11 in[JS87] cannot be applied, although $\mathcal{E}_n$ is a first entrance time. But with an analogous proof, and using the fact that $\mathcal{H}_n^{\mathrm{M}}$ is a Poisson process, we easily show that $\mathcal{E}_n \to \mathrm{e}$ a.s.

So, we have obtained the a.s. convergence (and thus the convergence in probability)

$$(\bar{\mathcal{Z}}_n(\mathcal{T}_n^x), \mathcal{L}_n(\mathcal{T}_n^x), \mathcal{E}_n) \xrightarrow[\mathbb{P}]{a.s.} (\bar{Z}(T^{-x}), L(T^{-x}), \mathrm{e})$$

which gives, together with the equality in law $(\bar{\mathcal{Z}}_n(\mathcal{T}_n^x), \mathcal{L}_n(\mathcal{T}_n^x), \mathcal{E}_n) \overset{\mathcal{L}}{=} (\bar{\tilde{Z}}_n(T_n^{-x}), L_n(T_n^{-x}), \mathrm{e}_n)$, the joint convergence in distribution of $(\bar{\tilde{Z}}_n(T_n^{-x}), L_n(T_n^{-x}), \mathrm{e}_n)$ towards $(Z(T^{-x}), L(T^{-x}), \mathrm{e})$. □

Lemma 3.18. *For all $y > 0$, $v > 0$ and $q \in \{0, 1\}$,*

$$\begin{aligned}
&\mathbb{P}_{-y}(Z(T^{(0,\infty)}) \in \mathrm{d}v,\ \Delta Z^{\mathrm{M}}(T^{(0,\infty)}) \in \mathrm{d}q) \\
&= \frac{b^2}{2}(W'(y) - \eta W(y))\delta_0(\mathrm{d}v)\delta_0(\mathrm{d}q) + \int_{(0,\infty)} (e^{-\eta u}W(y) - W(y-u))\ \mathbb{B}_{f(u+v)}(\mathrm{d}q)\ \Lambda(u + \mathrm{d}v)\ \mathrm{d}u,
\end{aligned}$$

and for all $n \geq 1$

$$\begin{aligned}
&\mathbb{P}_{-y}(\tilde{Z}_n(T_n^{(0,\infty)}) \in \mathrm{d}v,\ \Delta\tilde{Z}_n^{\mathrm{M}}(T_n^{(0,\infty)}) \in \mathrm{d}q) \\
&\qquad = \int_{(0,\infty)} (e^{-\tilde{\eta}_n v}\tilde{W}_n(y) - \tilde{W}_n(y-u))\ \mathbb{B}_{f_n(n(u+v))}(\mathrm{d}q)\ \tilde{\Lambda}_n(u + \mathrm{d}v)\ \mathrm{d}u.
\end{aligned}$$

The first quantity corresponds in fact to $g^x(y-x, \mathrm{d}q, \mathrm{d}v)$ introduced in the statement of Theorem 3.13. To keep consistency in the notation, we will then set

$$g_n^x(a, \mathrm{d}q, \mathrm{d}v) := \int_{(0,\infty)} (e^{-\tilde{\eta}_n v}\tilde{W}_n(x+a) - \tilde{W}_n(x+a-u))\mathbb{B}_{f_n(n(u+v))}(\mathrm{d}q)\tilde{\Lambda}_n(u + \mathrm{d}v)\mathrm{d}u,$$

which corresponds to the second formula of Lemma 3.18.

Proof :
We first write

$$\begin{aligned}
&\mathbb{P}_{-y}(\tilde{Z}_n(T_n^{(0,\infty)}) \in \mathrm{d}v,\ \Delta\tilde{Z}_n^{\mathrm{M}}(T_n^{(0,\infty)}) \in \mathrm{d}q) \\
&= \int_{[0,\infty)} \mathbb{P}_{-y}(\tilde{Z}_n(T_n^{(0,\infty)}) \in \mathrm{d}v,\ \tilde{Z}_n(T_n^{(0,\infty)}-) \in \mathrm{d}u)\mathbb{B}_{f_n(n(u+v))}(\mathrm{d}q).
\end{aligned}$$

and similarly for Z.
Now according to [Kyp06] (see consequence of (8.29)), we have for all $u > 0$, $v > 0$:

$$\mathbb{P}_{-y}(Z(T^{(0,\infty)}) \in \mathrm{d}v,\ Z(T^{(0,\infty)}-) \in \mathrm{d}u) = (e^{-\eta u}W(y) - W(y-u))\Lambda(u+\mathrm{d}v)\mathrm{d}u,$$

and similarly

$$\mathbb{P}_{-y}(\tilde{Z}_n(T_n^{(0,\infty)}) \in \mathrm{d}v,\ \tilde{Z}_n(T_n^{(0,\infty)}-) \in \mathrm{d}u) = (e^{-\tilde{\eta}_n v}\tilde{W}_n(y) - \tilde{W}_n(y-u))\tilde{\Lambda}_n(u+\mathrm{d}v)\mathrm{d}u.$$

Moreover, [Kyp06, Exercise 8.6] provides a formula for the probability of creeping over 0 starting at $-y < 0$ for a spectrally positive Lévy process, that is, the probability that the process is equal to 0 at $T^{(0,\infty)}$ under $\mathbb{P}_{-y}$. In particular this probability is zero if the process has no Gaussian component, so that at rank n we have

$$\begin{aligned}&\mathbb{P}_{-y}(\tilde{Z}_n(T_n^{(0,\infty)}) \in \mathrm{d}v,\ \Delta\tilde{Z}_n^{\mathrm{M}}(T_n^{(0,\infty)}) \in \mathrm{d}q)\\ &= \int_{(0,\infty)} (e^{-\tilde{\eta}_n v}\tilde{W}_n(y) - \tilde{W}_n(y-u))\tilde{\Lambda}_n(u+\mathrm{d}v)\,\mathrm{d}u\,\mathbb{B}_{f_n(n(u+v))}(\mathrm{d}q).\end{aligned}$$

On the other hand, as far as Z is concerned, its Gaussian coefficient $\frac{b^2}{2}$ might be positive, and since $f(0) = 0$, the formula of Exercise 8.6 in [Kyp06] :

$$\mathbb{P}_{-y}(Z(T^{(0,\infty)}) = 0) = \frac{b^2}{2}(W'(y) - \eta W(y))$$

implies

$$\begin{aligned}&\mathbb{P}_{-y}(Z(T^{(0,\infty)}) \in \mathrm{d}v,\ \Delta Z^{\mathrm{M}}(T^{(0,\infty)}) \in \mathrm{d}q)\\ &= \frac{b^2}{2}(W'(y) - \eta W(y))\delta_0(\mathrm{d}v)\delta_0(\mathrm{d}q) + \int_{(0,\infty)} (e^{-\eta u}W(y) - W(y-u))\Lambda(u+\mathrm{d}v)\,\mathrm{d}u\,\mathbb{B}_{f(u+v)}(\mathrm{d}q),\end{aligned}$$

which ends the proof. □

Lemma 3.19. *For all $n \geq 1$, $y \in \mathbb{R}_+$, $r \in \mathbb{R}_+$, $q \in \{0,1\}$, define*

$$\gamma_n^x(y,q,r) := \int_{(0,\infty)} e^{-rv} g_n^x(y,\{q\},\mathrm{d}v).$$

Then the Laplace transform $\gamma_n^x(y,q,r)$ converges towards $\gamma^x(y,q,r)$ (defined in Proposition 3.16) as $n \to \infty$, and the convergence is uniform w.r.t. y on every compact set of $\mathbb{R}_+$.

Proof :
Fix $y \in \mathbb{R}_+$, $r \in \mathbb{R}_+$, $q \in \{0,1\}$. Using the expression of g^x given by formula (13), we have :

$$\gamma_n^x(y,q,r) = \mathbb{E}_{-(x+y)}\big(e^{-r\tilde{Z}_n(T_n^{(0,\infty)})},\ \Delta\tilde{Z}_n^{\mathrm{M}}(T_n^{(0,\infty)}) = q\big),$$

which we also can reexpress as :

$$\begin{aligned}\gamma_n^x(y,q,r) &= \mathbb{E}\big(e^{-r(H_n^+(L_n(T_n^{(x+y,\infty)}))-(x+y))},\ \Delta H_n^{\mathrm{M}}(L_n(T_n^{(x+y,\infty)})) = q\big)\\ &= \int_{(x+y,\infty)\times(0,\infty)} \mathbb{B}_{f_n(nu)}(\mathrm{d}q)\, e^{-r(v-(x+y))}\,\mathbb{P}(H_n^+(L_n(T_n^{(x+y,\infty)})) \in \mathrm{d}v, \Delta\tilde{Z}_n(T_n^{(x+y,\infty)}) \in \mathrm{d}u).\end{aligned}$$

In the same vein, we have

$$\gamma^x(y,q,r) = \int_{(x+y,\infty)\times[0,\infty)} \mathbb{B}_{f(u)}(\mathrm{d}q)\, e^{-r(v-(x+y))}\, \mathbb{P}(H^+(L(T^{(x+y,\infty)})) \in \mathrm{d}v, \Delta Z(T^{(x+y,\infty)}) \in \mathrm{d}u).$$

To start with, we prove that the measures $\mathbb{P}(H_n^+(L_n(T_n^{(x+y,\infty)})) \in \mathrm{d}v,\ \Delta\tilde{Z}_n(T_n^{(x+y,\infty)}) \in \mathrm{d}u)$ converge weakly towards $\mathbb{P}(H^+(L(T^{(x+y,\infty)})) \in \mathrm{d}v,\ \Delta Z(T^{(x+y,\infty)}) \in \mathrm{d}u)$. First recall that thanks to Theorem 3.6 we have the convergence in distribution of $(H_n^+, \tilde{Z}_n)$ towards (H^+, Z). With probability one, we have that $T^{(x+y,\infty)}$ is either a continuity point of Z a.s., or it satisfies $Z(T^{(x+y,\infty)}-) < x+y < Z(T^{(x+y,\infty)})$ and $H^+(L(T^{(x+y,\infty)})-) < x+y < H^+(L(T^{(x+y,\infty)}))$ a.s. Note furthermore that the first entrance time of H^+ in $(x+y,\infty)$ is equal to $L(T^{(x+y,\infty)})$ a.s. Then we can easily adapt the proof of Proposition VI.2.12 in [JS87] to get that

$$(H_n^+(L_n(T_n^{(x+y,\infty)})),\ \Delta\tilde{Z}_n(T_n^{(x+y,\infty)})) \Rightarrow (H^+(L(T^{(x+y,\infty)})),\ \Delta Z(T^{(x+y,\infty)})).$$

On the other hand, under B.1 as well as under B.2, we have the uniform convergence of $(u,v) \mapsto \mathbb{B}_{f_n(nu)}\{q\}\, e^{-rv}$ to $(u,v) \mapsto \mathbb{B}_{f(u)}\{q\}\, e^{-rv}$ on every compact set of $(0,\infty)\times(x+y,\infty)$. Then from an appeal to Lemma A.1 we get the convergence of $\gamma_n^x(y,q,r)$ towards $\gamma^x(y,q,r)$ for all fixed $y, r \geq 0$ and $q \in \{0,1\}$.

It remains to prove the uniform convergence of γ_n^x w.r.t. the first variable, y, on every compact set of $\mathbb{R}_+$.
Take $r \geq 0$ and $q \in \{0,1\}$. For all $y \geq 0$, $\tilde{W}_n(x+y)$ is positive, and we set $\tilde{\gamma}_n^x(y,q,r) = \gamma_n^x(y,q,r)/\tilde{W}_n(x+y)$. Observe that $y \mapsto \tilde{\gamma}_n^x(y,q,r)$ is decreasing on $\mathbb{R}_+$: Indeed, it can be shown with elementary calculations that

$$\tilde{\gamma}_n^x(y,q,r) = \int_{(0,\infty)} \tilde{\Lambda}_n(\mathrm{d}z)\mathbb{B}_{f_n(nz)}(\{q\}) \int_0^z e^{-r(z-u)}\, e^{-\tilde{\eta}_n u} N_n(\inf\epsilon \leq -(x+y) \,|\, -\epsilon(\zeta-) = u)\, \mathrm{d}u,$$

and since the mappings $y \mapsto N_n(\inf\epsilon \leq -(x+y) \,|\, -\epsilon(\zeta-) = u)$ are clearly decreasing, we have the same property for $\tilde{\gamma}_n^x(\cdot,q,r)$. Next, recalling that the functions $\tilde{W}_n(x+\cdot)$ are strictly increasing and take positive values, we get that the functions $\tilde{\gamma}_n^x(\cdot,q,r)$ are decreasing, which leads to the uniform convergence of $\gamma_n^x(\cdot,q,r)$ to $\gamma^x(\cdot,q,r)$ on every compact set of $\mathbb{R}_+$. □

In the proof of Proposition 3.16, we will make a frequent use of the Markov property, applied alternately to Z (resp. $\tilde{Z}_n$) or to (H^+, H^{M}) (resp. $(H_n^+, H_n^{\mathrm{M}})$), at different stopping times. We already know that T^{-x}, $L^{-1}(t)$ (resp. T_n^{-x}, L_n^{-1}) are $\mathcal{F}$- (resp. $\mathcal{F}_n$-) stopping times. We introduce here three other stopping times which we will need later.

First we define the processes $\hat{Z}_n^{\mathrm{M}}, \hat{Z}^{\mathrm{M}}$ as follows : for all $t \geq 0$,

$$\hat{Z}_n^{\mathrm{M}}(t) := H_n^{\mathrm{M}}(L_n(t)-) \ \text{ and } \ \hat{Z}^{\mathrm{M}}(t) = H^{\mathrm{M}}(L(t)-).$$

The process $\hat{Z}_n^{\mathrm{M}}$ is a counting process which jumps every time a mutation occurs at a record time of $\tilde{Z}_n$: it can be seen as the matching process of H_n^{M} in the real time scale (in opposition to the local time scale) - and similarly for $\hat{Z}^{\mathrm{M}}$. We then have the identity $H_n^{\mathrm{M}} = \hat{Z}_n^{\mathrm{M}} \circ L_n^{-1}$.
We enlarge the initially considered filtrations and set for all $t \geq 0$:

$$\mathcal{F}_t^{\mathrm{M}} := \sigma(Z(s), \hat{Z}^{\mathrm{M}}(s),\ s \leq t)$$

and

$$\mathcal{F}_{n,t}^{\mathrm{M}} := \sigma(\tilde{Z}_n(s), \hat{Z}_n^{\mathrm{M}}(s), L_n(s)\, s \leq t).$$

We denote by $\mathcal{F}^{\mathrm{M}}$ (resp. $\mathcal{F}^{\mathrm{M}}_n$) the filtration $(\mathcal{F}^{\mathrm{M}}_t)_{t\geq 0}$ (resp. $(\mathcal{F}^{\mathrm{M}}_{n,t})_{t\geq 0}$). Finally, the notations $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$, $\mathcal{F}^{\mathrm{M}}_{n,L_n^{-1}}$ will respectively stand for the filtrations $(\mathcal{F}^{\mathrm{M}}_{L^{-1}(t)})_{t\geq 0}$ and $(\mathcal{F}^{\mathrm{M}}_{n,L_n^{-1}(t)})_{t\geq 0}$. Note that $(Z,\hat{Z}^{\mathrm{M}})$ is not a Markov process in the filtration $\mathcal{F}^{\mathrm{M}}$.

Proposition 3.20. (i) e *(resp.* e_n*) is a stopping time w.r.t. the filtration* $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$ *(resp.* $\mathcal{F}^{\mathrm{M}}_{n,L_n^{-1}}$*).*

(ii) $L^{-1}(\mathrm{e}-)$ *(resp.* $L_n^{-1}(\mathrm{e}_n-)$*) is a stopping time w.r.t. the filtration* $\mathcal{F}^{\mathrm{M}}$ *(resp.* $\mathcal{F}^{\mathrm{M}}_n$*).*

(iii) $L(T^{-x})$ *(resp.* $L_n(T_n^{-x})$*) is a stopping time w.r.t. the filtration* $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$ *(resp.* $\mathcal{F}^{\mathrm{M}}_{n,L_n^{-1}}$*).*

Proof :

(i) e is the first entrance time of the bivariate subordinator (H^+,H^{M}) into $\mathbb{R}_+\times\mathbb{R}_+^*$. Thus e is a stopping time w.r.t. the natural filtration associated to (H^+,H^{M}), and then w.r.t. $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$.

(ii) We have for all $t\geq 0$:

$$\{L^{-1}(\mathrm{e}-)\leq t\}=\bigcap_{u\in\mathbb{Q}\cap\mathbb{R}_+^*}\big(\{L^{-1}(u)\leq t\}\cap\{u<\mathrm{e}\}\big),$$

now $L^{-1}(t)$ is a $\mathcal{F}$- stopping time, thus $\{L^{-1}(u)\leq t\}\in\mathcal{F}_t$, and e is a $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$-stopping time, thus $\{\mathrm{e}>u\}\in\mathcal{F}^{\mathrm{M}}_{L^{-1}(u)}$. Consequently $\{L^{-1}(u)\leq t\}\cap\{u<\mathrm{e}\}$ belongs to $\mathcal{F}^{\mathrm{M}}_t$ for all $u>0$, and so is $\{L^{-1}(\mathrm{e}-)\leq t\}$.

(iii) For all $t\geq 0$, we want to prove that $\{L(T^{-x})\leq t\}=\{T^{-x}\leq L^{-1}(t)\}$ a.s. For a clearer view of what follows, see Figure 6.
Fix $u\geq 0$. On the one hand, since $u\leq L^{-1}(L(u))$ and L^{-1} is increasing, $L(u)\leq t$ implies $u\leq L^{-1}(t)$. On the other hand, in the infinite variation case, the function $L\circ L^{-1}$ is the identity function, and hence $u\leq L^{-1}(t)$ implies $L(u)\leq t$. In the finite variation case, the definition of L^{-1} implies that if $u<L^{-1}(t)$, then $L(u)\leq t$. Now the event $\{\exists t\geq 0,\ T^{-x}=L^{-1}(t)\}$ is negligible, thus $\{T^{-x}\leq L^{-1}(t)\}=\{T^{-x}<L^{-1}(t)\}$ a.s.
We conclude from what precedes that the events $\{L(T^{-x})\leq t\}$ and $\{T^{-x}\leq L^{-1}(t)\}$ are identical a.s., and since T^{-x} is a stopping time w.r.t. the filtration $\mathcal{F}$, this implies that $L(T^{-x})$ is a stopping time w.r.t. the filtration $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$.

Remark that the three proofs above work in the infinite variation case as well as in the finite variation case, so that the conclusions are also true for e_n, $L_n^{-1}(\mathrm{e}_n-)$, $L_n(T_n^{-x})$, $n\geq 1$.

□

Proof of Proposition 3.16 :
We begin with the computation of the probability measure $\mathbb{P}(H^+(\mathrm{e}-)\in\mathrm{d}z,\ L(T^{-x})<\mathrm{e})$: The calculation below is done for the limiting process. However we pay attention to the fact that the arguments are still valid in the finite variation case, so that the same calculation remains true for $\mathbb{P}(H_n^+(\mathrm{e}_n-)\in\mathrm{d}z,\ L_n(T_n^{-x})<\mathrm{e}_n)$.

Noting that $L(T^{-x})<\mathrm{e}$ coincides with $H^{\mathrm{M}}(L(T^{-x}))=0$ a.s., and then applying the Markov property to the process (H^+,H^{M}) at the $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$-stopping time $L(T^{-x})$, we have

$$\begin{aligned}&\mathbb{P}(H^+(\mathrm{e}-)\in\mathrm{d}z,\ L(T^{-x})<\mathrm{e})\\&=\int_{[0,z)}\int_{(a,z)}\mathbb{P}_b(H^+(\mathrm{e}-)\in\mathrm{d}z)\,\mathbb{P}(H^+(L(T^{-x})-)\in\mathrm{d}a,\ H^+(L(T^{-x}))\in\mathrm{d}b,\ H^{\mathrm{M}}(L(T^{-x}))=0)\end{aligned}$$

Using the notation $d(T^{-x}) := L^{-1}(L(T^{-x}))$, recall that

$$H^+(L(T^{-x})-) = \bar{Z}(T^{-x}) \text{ and } H^+(L(T^{-x})) = \bar{Z}(d(T^{-x})).$$

Furthermore, with probability one

$$\{H^{\mathrm{M}}(L(T^{-x})) = 0\} = \{\hat{Z}^{\mathrm{M}}(T^{-x}) = 0\} \cap \{\Delta\hat{Z}^{\mathrm{M}}(d(T^{-x})) = 0\}.$$

Conditional on $\bar{Z}(T^{-x})$, the random variable $\Delta\hat{Z}^{\mathrm{M}}(d(T^{-x}))$ is independent from $\mathcal{F}^{\mathrm{M}}_{T^{-x}}$, and has the law of $\Delta Z^{\mathrm{M}}(T^{(0,\infty)})$ under $\mathbb{P}_{-x-\bar{Z}(T^{-x})}$ (Note that $T^{(0,\infty)}$ is a.s. necessarily a record time for Z under $\mathbb{P}_{-x-\bar{Z}(T^{-x})}$). We then use the Markov property again, applied to the process Z at the $\mathcal{F}$-stopping time T^{-x} :

$$\begin{aligned}
&\mathbb{P}(H^+(L(T^{-x})-) \in \mathrm{d}a,\ H^+(L(T^{-x})) \in \mathrm{d}b,\ H^{\mathrm{M}}(L(T^{-x})) = 0)\\
&\quad = \mathbb{P}(\bar{Z}(T^{-x}) \in \mathrm{d}a,\ \hat{Z}^{\mathrm{M}}(T^{-x}) = 0)\ \mathbb{P}_{-(a+x)}(Z(T^{(0,\infty)}) \in \mathrm{d}b - a,\ \Delta Z^{\mathrm{M}}(T^{(0,\infty)}) = 0)\\
&\quad = \mathbb{P}(\bar{Z}(T^{-x}) \in \mathrm{d}a,\ L(T^{-x}) \le \mathrm{e})\ \mathbb{P}_{-(a+x)}(Z(T^{(0,\infty)}) \in \mathrm{d}b - a,\ \Delta Z^{\mathrm{M}}(T^{(0,\infty)}) = 0),
\end{aligned}$$

and finally, using the notation introduced in the statement of Theorem 3.13 and Lemma 3.18, this gives

$$\begin{aligned}
&\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L(T^{-x}) < \mathrm{e})\\
&\qquad = \mathbb{P}(\Delta H^+(\mathrm{e}) \in \mathrm{d}y) \int_{[0,z)} \int_{(a,z)} \mathbb{P}_b(H^+(\mathrm{e}-) \in \mathrm{d}z)\ \pi(\mathrm{d}a)\ g^x(a, \{0\}, \mathrm{d}b - a).
\end{aligned}$$

When Z does not drift to $-\infty$, by definition of e as first entrance time and thanks to Proposition 0.5.2 in [Ber96], we have

$$\mathbb{P}_b(H^+(\mathrm{e}-) \in \mathrm{d}z) = \mathbb{P}_b(H^*(\alpha) \in \mathrm{d}z),$$

where H^* is the subordinator defined in Theorem 3.4, and α is an independent exponential random variable with parameter λ.
In the same way, we treat the case Z drifts to $-\infty$ appealing to [Ber96, Prop.0.5.2] and to Theorem 3.3 : set $\varrho := \mathrm{e} \wedge \mathcal{K}$, where $\mathcal{K}$ is an independent exponential variable with parameter k. Then ϱ follows an exponential distribution with parameter $\lambda + k$, and we have

$$\mathbb{P}_b(H^+(\mathrm{e}-) \in \mathrm{d}z) = \frac{\lambda}{\lambda+k}\mathbb{P}_b(H^+(\varrho-) \in \mathrm{d}z \,|\, \varrho = \mathrm{e}) = \frac{\lambda}{\lambda+k}\mathbb{P}_b(H^*(\alpha') \in \mathrm{d}z),$$

where α' is an independent exponential random variable with parameter $\lambda + k$.

By definition of $U^{(\cdot)}_*$ we have then in both cases $\mathbb{P}_b(H^+(\mathrm{e}-) \in \mathrm{d}z) = \lambda U^{(\lambda+k)}_*(\mathrm{d}z - b)$ (recall that we set $k = 0$ if Z does not drift to $-\infty$). As a consequence,

$$\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ L(T^{-x}) < \mathrm{e}) = \lambda \int_{[0,z)} \int_{(a,z)} U^{(\lambda+k)}_*(\mathrm{d}z - b)\ \pi(\mathrm{d}a)\ g^x(a, \{0\}, \mathrm{d}b - a).$$

Hence we get for the Laplace transform :

$$\mathbb{E}(e^{-rH^+(\mathrm{e}-)},\ L(T^{-x}) < \mathrm{e}) = \int_{[0,\infty)} \pi(\mathrm{d}a) \int_{[a,\infty)} g^x(a, \{0\}, \mathrm{d}b - a) \int_{[b,\infty)} e^{-rz}\lambda U^{(\lambda+k)}_*(\mathrm{d}z - b).$$

From the definition of $U_*^{(\lambda+k)}$ we have for all $r \geq 0$, $\int_{(0,\infty)} e^{-rz} U_*^{(\lambda+k)}(\mathrm{d}z) = (\lambda + k + \psi^*(r))^{-1}$, which leads to

$$\begin{aligned}\mathbb{E}(e^{-rH^+(\mathrm{e}-)},\ L(T^{-x}) < \mathrm{e}) &= \frac{\lambda}{\lambda + k + \psi^*(r)} \int_{[0,\infty)} \pi(\mathrm{d}a) \int_{[a,\infty)} g^x(a, \{0\}, \mathrm{d}b - a) e^{-br} \\ &= \frac{\lambda}{\lambda + k + \psi^*(r)} \int_{[0,\infty)} \pi(\mathrm{d}a) \gamma^x(a, 0, r) e^{-ar},\end{aligned}$$

and as announced, we have a similar formula at rank n :

$$\mathbb{E}(e^{-rH_n^+(\mathrm{e}_n-)},\ L_n(T_n^{-x}) < \mathrm{e}_n) = \frac{\lambda_n}{\lambda_n + k_n + \psi_n^*(r)} \int_{[0,\infty)} \pi_n(\mathrm{d}a) \gamma_n^x(a, 0, r) e^{-ar}.$$

Now as $n \to \infty$, thanks to [Del13a, Prop.4.9.(i)] $\lambda_n = \mu_n(\mathbb{R}_+^*, \{1\})$ converges to λ, and thanks to Theorem 3.4 ψ_n^* converges to ψ^*. According to the proof of 3.3 in Chapter I, we also have $k_n \to k$. As for the integral, thanks to Lemma 3.17 and Lemma 3.19 we can apply Lemma A.1, and hence we have proved that as $n \to \infty$,

$$\mathbb{E}(e^{-rH_n^+(\mathrm{e}_n-)},\ L_n(T_n^{-x}) < \mathrm{e}_n) \to \mathbb{E}(e^{-rH^+(\mathrm{e}-)},\ L(T^{-x}) < \mathrm{e})$$

for all $r \geq 0$. This finishes the proof. □

Finally, before we prove the theorem, we need the following technical lemma :

Lemma 3.21. *The event $\{L(T^{-x}) < \mathrm{e}\}$ (resp. $\{L_n(T_n^{-x}) < \mathrm{e}_n\}$) belongs to $\mathcal{F}^{\mathrm{M}}_{L^{-1}(\mathrm{e}-)}$(resp. $\mathcal{F}^{\mathrm{M}}_{n,L_n^{-1}(\mathrm{e}_n-)}$).*

Proof :
We first want to prove that $\{L(T^{-x}) < \mathrm{e}\} = \{L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e})\}$ a.s. (and the equivalent equality at rank n).

As far as the limiting process is concerned, we are in the infinite variation case : The process L^{-1} is a.s. continuous and strictly increasing, so that $\{L(T^{-x}) < \mathrm{e}\} = \{L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e})\}$ a.s. In fact these two events still coincide a.s. in the finite variation case, although L^{-1} is not strictly increasing : Indeed, L^{-1} is injective on the set of all jumping times of H^+ ; now $L(T^{-x})$ and e are a.s. two jumping times of H^+, hence $L^{-1}(L(T^{-x})) = L^{-1}(\mathrm{e})$ implies $L(T^{-x}) = \mathrm{e}$, and the claim is proved.

We now prove that $L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e}) \Leftrightarrow T^{-x} \leq L^{-1}(\mathrm{e}-)$ a.s.
On the one hand, $L^{-1}(L(T^{-x}))$, $L^{-1}(\mathrm{e})$ and $L^{-1}(\mathrm{e}-)$ belong to the zero set of $\bar{Z} - Z$, and $L^{-1}(\mathrm{e})$ and $L^{-1}(\mathrm{e}-)$ are two consecutive (possibly equal) zeros of $\bar{Z} - Z$. Thus $L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e})$ implies $L^{-1}(L(T^{-x})) \leq L^{-1}(\mathrm{e}-)$, and since $T^{-x} \leq L^{-1}(L(T^{-x}))$ a.s., this ensures that with probability one $\{L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e})\} \subset \{T^{-x} \leq L^{-1}(\mathrm{e}-)\}$.
On the other hand, assume that $T^{-x} \leq L^{-1}(\mathrm{e}-)$. The event $\{T^{-x} = L^{-1}(t),\ \text{for some } t \geq 0\}$ is negligible and thus by definition of $L^{-1}(\mathrm{e}-)$, there exists $u < \mathrm{e}$ such that $T^{-x} < L^{-1}(u)$ a.s. This ensures that $L^{-1}(L(T^{-x})) = \inf\{L^{-1}(u),\ L^{-1}(u) > T^{-x}\} < L^{-1}(\mathrm{e})$ a.s., and then $\{T^{-x} \leq L^{-1}(\mathrm{e}-)\} \subset \{L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e})\}$

So, we have proved that almost surely $\{L(T^{-x}) < \mathrm{e}\} = \{T^{-x} \leq L^{-1}(\mathrm{e}-)\}$ a.s. We conclude using the fact that T^{-x} is a $\mathcal{F}^{\mathrm{M}}$-stopping time. The proof above remains true in the finite variation case, so that the result is also valid at rank n. □

Proof of Theorem 3.13 :
To begin with, we prove formulas (12) and (14). As in the proof above, we do the calculation and reasoning for the limiting process Z, and we add some remarks when needed so that it remains valid at rank n.

First note that thanks to the Markov property applied to (H^+, H^{M}) at the $\mathcal{F}^{\mathrm{M}}_{L^{-1}}$-stopping time e, and since the event $\{L^{-1}(\mathrm{e}) < T^{-x}\} = \{T^{-x} \leq L^{-1}(\mathrm{e})\}^c$ (where A^c denotes the complementary event of A) belongs to $\mathcal{F}^{\mathrm{M}}_{L^{-1}(\mathrm{e})}$, we have

$$\begin{aligned}&\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L^{-1}(\mathrm{e}) < T^{-x} < T^{(\tau-x,\infty)})\\ &= \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L^{-1}(\mathrm{e}) < T^{-x})\mathbb{P}_{z+y}(T^{-x} < T^{(\tau-x,\infty)}),\end{aligned}$$

where $\mathbb{P}_{z+y}(T^{-x} < T^{(\tau-x,\infty)}) = W(\tau - x - z - y)/W(\tau)$. Now we have :

$$\begin{aligned}&\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\, \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L^{-1}(\mathrm{e}) < T^{-x})\\ &= \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\, \Delta H^+(\mathrm{e}) \in \mathrm{d}y) - \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y, L(T^{-x}) \leq \mathrm{e})\\ &= \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\, \Delta H^+(\mathrm{e}) \in \mathrm{d}y) - \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y, L(T^{-x}) < \mathrm{e}),\\ &\qquad\qquad - \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L(T^{-x}) = \mathrm{e}),\end{aligned}$$

where in the last equality we distinguished the case where the first mutation, in the time scale of Z, occurs at the end of the excursion interval containing T^{-x}, or later. Recall that despite the fact that L^{-1} shall not be strictly increasing (finite variation case), we always have $L(T^{-x}) < \mathrm{e} \Leftrightarrow L^{-1}(L(T^{-x})) < L^{-1}(\mathrm{e})$ (see proof of Lemma 3.21).

As in the proof of Proposition 3.16, applying Proposition 0.5.2 in [Ber96], we get for the first term in the sum :

$$\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z, \Delta H^+(\mathrm{e}) \in \mathrm{d}y) = \lambda U_*^{(\lambda+k)}(\mathrm{d}z)\mathbb{P}(\Delta H^+(\mathrm{e}) \in \mathrm{d}y),$$

Then we compute the second term in the sum : Lemma 3.21 ensures that $\{L(T^{-x}) < \mathrm{e}\} \in \mathcal{F}^{\mathrm{M}}_{L^{-1}(\mathrm{e}-)}$ a.s., thus by Markov property applied to Z at the $\mathcal{F}^{\mathrm{M}}$-stopping time $L^{-1}(\mathrm{e}-)$ we have

$$\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\, \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\, L(T^{-x}) < \mathrm{e}) = \mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\, L(T^{-x}) < \mathrm{e})\mathbb{P}(\Delta H^+(\mathrm{e}) \in \mathrm{d}y),$$

and thanks to the calculation made in the proof of Proposition 3.16, we get

$$\begin{aligned}&\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L(T^{-x}) < \mathrm{e}) =\\ &\qquad \lambda\mathbb{P}(\Delta H^+(\mathrm{e}) \in \mathrm{d}y) \int_{[0,z)}\int_{(a,z)} U_*^{(\lambda+k)}(\mathrm{d}z - b)\ \pi(\mathrm{d}a)\ g^x(a, \{0\}, \mathrm{d}b - a).\end{aligned}$$

Finally note that $\lambda\mathbb{P}(\Delta H^+(\mathrm{e}) \in \mathrm{d}y) = \theta\delta_0(\mathrm{d}y)$ under Assumption B.1, $\lambda\mathbb{P}(\Delta H^+(\mathrm{e}) \in \mathrm{d}y) = \mu(\mathrm{d}y, \{1\}) + \rho\delta_0(\mathrm{d}y)$ under Assumption B.2, and $\lambda_n\mathbb{P}(\Delta H_n^+(\mathrm{e}_n) \in \mathrm{d}y) = \mu_n(\mathrm{d}y, \{1\})$ in both cases.

It remains to compute the third term in the sum. With an application of the Markov property to Z at T^{-x} as in the proof of Proposition 3.16, we get

$$\mathbb{P}(H^+(\mathrm{e}-) \in \mathrm{d}z,\ \Delta H^+(\mathrm{e}) \in \mathrm{d}y,\ L(T^{-x}) = \mathrm{e}) = \pi(\mathrm{d}z)g^x(z, \{1\}, \mathrm{d}y),$$

which vanishes in case B.1 according to Remark 3.14.

Thereby we have established formulas (12) and (14), and these formulas remain true at rank n (considering in case B.2 that the coefficient ρ is zero in the finite variation case). Since the expression of g^x given by formula (13) in the statement of the theorem has been established in Lemma 3.18, proving the claimed convergence will end the proof.

From the calculation above we have at rank n (in both cases B.1 and B.2) :

$$\begin{aligned}\mathbb{P}(H_n^+(\mathrm{e}_n-) \in \mathrm{d}z,\ \Delta H_n^+(\mathrm{e}_n) \in \mathrm{d}y,\ L_n^{-1}(\mathrm{e}_n) < T_n^{-x} < T_n^{(\tau-x,\infty)}) \\ = \frac{\tilde{W}_n(\tau-z-y)}{\tilde{W}_n(\tau)}\Bigg\{\mu_n(\mathrm{d}y,\{1\})\Big[U_{n,*}^{(\lambda_n+k_n)}(\mathrm{d}z) - \mathbb{P}(H_n^+(\mathrm{e}_n-) \in \mathrm{d}z,\ L_n(T_n^{-x}) < \mathrm{e}_n)\Big] \\ - \pi_n(\mathrm{d}z)g_n^x(z,\{1\},\mathrm{d}y)\Bigg\}, \quad (15)\end{aligned}$$

where π_n, g_n^x have been defined respectively in Lemmas 3.17 and 3.19, and $U_{n,*}^{(\lambda_n+k_n)}$ denotes the (λ_n+k_n)-resolvent measure of H_n^* (defined in Theorem 3.4). Now as $n \to \infty$, for all $z \geq 0$, $y > 0$,

- From Proposition 3.1.(iii), we know that $\tilde{W}_n(\tau-x-z-y)/\tilde{W}_n(\tau)$ converges to $W(\tau-x-z-y)/W(\tau)$.
- From [Del13a, Prop.4.9.(i)], $\lambda_n = \mu_n(\mathbb{R}_+^*,\{1\})$ converges to λ, and $\mu_n(\mathrm{d}y,\{1\})$ converges weakly to $\theta\delta_0(\mathrm{d}y)$ (resp. $\mu(\mathrm{d}y,\{1\}) + \rho\delta_0(\mathrm{d}y)$) in case B.1 (resp. B.2).
- The Laplace transform of the measure $U_{n,*}^{(\lambda_n+k_n)}$ (resp. $U_*^{(\lambda+k)}$) is given by $(\lambda_n+k_n+\psi_n^*(\cdot))^{-1}$ (resp. $(\lambda+k+\psi^*(\cdot))^{-1}$), hence the measure $U_{n,*}^{(\lambda_n+k_n)}$ converges weakly towards $U_*^{(\lambda+k)}$ using Theorem 3.4 and the fact that $k_n \to k$ (see proof of Theorem 3.3 in Chapter I).
- The weak convergence of the probability measure $\mathbb{P}(H_n^+(\mathrm{e}_n) \in \mathrm{d}z,\ L_n(T_n^{-x}) < \mathrm{e}_n)$ has been proved via the convergence of its Laplace transform in Proposition 3.16.
- Finally, the weak convergence of $\pi_n(\mathrm{d}z)g_n^x(z,\{1\},\mathrm{d}y)$ to $\pi(\mathrm{d}z)g^x(z,\{1\},\mathrm{d}y)$ is straightforward from Lemmas 3.17 and 3.19.

As a conclusion we have proved the weak convergence under Assumption B.2 (resp. B.1) of (15) to (12) (resp. (14)) . □

3.3.2 Convergence in distribution of $(M_{n,\varepsilon})_n$ towards M_ε

The aim of this last subsection is to prove Proposition 3.9, appealing to Theorem 1 in [Kar75]. The four lemmas below ensure that the conditions needed to apply this theorem are fulfilled : First in Lemma 3.22, we make use of Theorem 3.13 to obtain a slightly more precise result about the convergence of the transition measures ν^{M} and ν^{D}. Second, we give in Lemma 3.23 explicit expressions for $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ and $\nu_\varepsilon^{\mathrm{INIT}}$, which allow us to prove in Lemma 3.24 the weak convergence of $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ towards $\nu_\varepsilon^{\mathrm{INIT}}$. Finally we deduce from this the convergence in distribution of $M_{n,\varepsilon}(0)$ towards $M_\varepsilon(0)$.

Lemma 3.22. *Suppose $x_n \to x$ as $n \to \infty$, where x_n and x are positive real numbers. Then $\nu_n^{\mathrm{M}}(x_n,\cdot)$ (resp. $\nu_n^{\mathrm{D}}(x_n,\cdot)$) converges weakly towards $\nu^{\mathrm{M}}(x,\cdot)$ (resp. $\nu^{\mathrm{D}}(x,\cdot)$).*

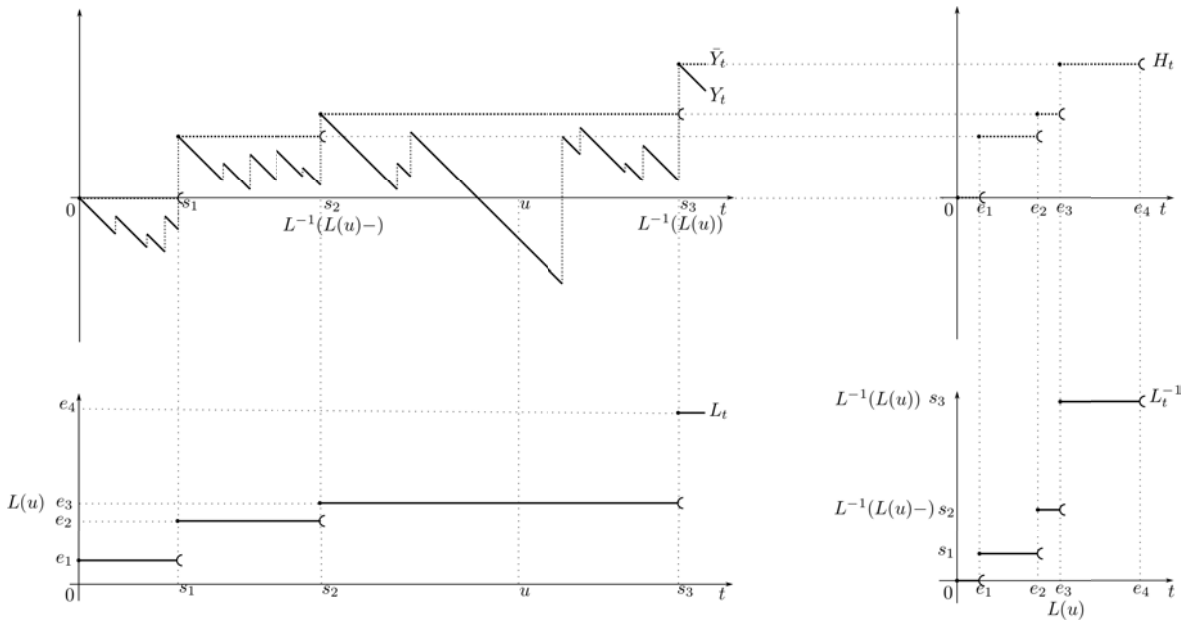

FIGURE 6 – An example of the paths of Z, its local time at the supremum and its ladder process in the finite variation case.

Proof :
Let A be a set in $\mathcal{B}([\varepsilon,\tau)\times\{0,1\})$ satisfying $\nu^{\mathrm{M}}(x,\partial A)=0$ (where ∂A denotes the boundary of A). First write

$$|\nu_n^{\mathrm{M}}(x_n,A)-\nu^{\mathrm{M}}(x,A)|\le|\nu_n^{\mathrm{M}}(x_n,A)-\nu_n^{\mathrm{M}}(x,A)|+|\nu_n^{\mathrm{M}}(x,A)-\nu^{\mathrm{M}}(x,A)|.$$

The second term in the right-hand side vanishes thanks to Theorem 3.13. Besides, we have

$$|\mathbb{P}_0(H_n^+(\mathrm{e}_n)\in A, L_n^{-1}(\mathrm{e}_n)<T_n^{-x_n})-\mathbb{P}_0(H_n^+(\mathrm{e}_n)\in A, L_n^{-1}(\mathrm{e}_n)<T^{-x})|\le\mathbb{P}_0(T_n^{-x}<T_n^{-x_n}),$$

which vanishes as $n\to\infty$ thanks to the a.s. continuity of $x\mapsto T_n^{-x}$ on $\mathbb{R}_+$ under $\mathbb{P}_0$. Then, by definition of ν_n^{M} (see (4)), we get $|\nu_n^{\mathrm{M}}(x_n,A)-\nu_n^{\mathrm{M}}(x,A)|\to 0$ as $n\to\infty$ (for the sake of simplicity, we omitted here the conditioning that appears in the definition of ν_n^{M}).
A similar reasoning holds for ν^{D} (the weak convergence of ν_n^{D} towards ν^{D} is a consequence of Lemma 3.17). □

Lemma 3.23. *For all $(u,q)\in[\varepsilon,\tau)\times\{0,1\}$, we have*

$$\begin{aligned}&\nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}u,\mathrm{d}q)\\&=\left(\frac{1}{\tilde{W}_n(\varepsilon)}-\frac{1}{\tilde{W}_n(\tau)}\right)\frac{\tilde{W}_n(\tau-u)}{\tilde{W}_n(\tau)}\mathrm{d}u\int_{(u,\infty)}\tilde{\Lambda}_n(\mathrm{d}z)\mathbb{B}_{f_n(nz)}(\mathrm{d}q)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right),\end{aligned}\tag{16}$$

$$\begin{aligned}&\nu_{\varepsilon}^{\mathrm{INIT}}(\mathrm{d}u,\mathrm{d}q)\\&=\frac{1}{p_\varepsilon}\frac{W(\tau-u)}{W(\tau)}\left[\frac{b^2}{2}\frac{W'(\varepsilon)}{W(\varepsilon)}\delta_\varepsilon(\mathrm{d}u)\delta_0(\mathrm{d}q)+\mathrm{d}u\int_{(u,\infty)}\Lambda(\mathrm{d}z)\mathbb{B}_{f(z)}(\mathrm{d}q)\left(1-\frac{W(\varepsilon-(z-u))}{W(\varepsilon)}\right)\right],\end{aligned}\tag{17}$$

Lemma 3.24. *The sequence of measures* $(\nu_{n,\varepsilon}^{\text{INIT}})$ *converges weakly towards* $\nu_{\varepsilon}^{\text{INIT}}$.

For the sake of clarity we only prove the two lemmas for $\nu_{n,\varepsilon}^{\text{INIT}}(\,\cdot\,,\{0,1\})$ and $\nu_{\varepsilon}^{\text{INIT}}(\,\cdot\,,\{0,1\})$.

Proof of Lemma 3.23 :
We begin by proving (16). Recall the following definition :

$$\nu_{n,\varepsilon}^{\text{INIT}}(\,\cdot\,,\{0,1\}) = \frac{1}{{p'}_{n,\varepsilon}}\mathbb{P}_0(\Upsilon_n \in \,\cdot\,,\ T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}).$$

Applying the strong Markov property at $T_n^{-\varepsilon}$, we get :

$$\mathbb{P}_0(\Upsilon_n - \varepsilon \in \mathrm{d}u,\ T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}) = \mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)})\mathbb{P}_{-\varepsilon}(\Upsilon_n - \varepsilon \in \mathrm{d}u,\ T_n^{(0,\infty)} < T_n^{-\tau}).$$

Now conditional on $T_n^{(0,\infty)} < T_n^{-\tau}$, $\Upsilon_n - \varepsilon$ has the distribution under $N_n'(\cdot \mid -\inf\epsilon < \tau - \varepsilon,\ \sup\epsilon \geq \varepsilon)$ of the undershoot of an excursion at its first entrance time in $(0,\infty)$. Thanks to Proposition 0.5.2(ii) in [Ber96], we then have

$$\begin{aligned}
&\mathbb{P}_0(\Upsilon_n - \varepsilon \in \mathrm{d}u,\ T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}) \\
&\qquad = \mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)})\mathbb{P}_{-\varepsilon}(T_n^{(0,\infty)} < T_n^{-\tau})\frac{N_n'(-\epsilon(\chi-) \in \mathrm{d}u,\ -\inf\epsilon < \tau - \varepsilon,\ \sup\epsilon \geq \varepsilon)}{N_n'(-\inf\epsilon < \tau - \varepsilon, \sup\epsilon \geq \varepsilon)},
\end{aligned}$$

Recall that for any $\epsilon \in \mathscr{E}'$, $\chi(\epsilon)$ denotes the first (and unique) entrance time of ϵ into $(0,\infty)$, which is a.s. finite on $\{-\inf\epsilon < \tau - \varepsilon\}$.
The process $\tilde{Z}_n$ has finite variation, and it can be shown with elementary calculations that

$$N_n'(-\inf\epsilon < \tau - \varepsilon, \sup\epsilon \geq \varepsilon) = \frac{\tilde{W}_n(0)}{\tilde{W}_n(\tau-\varepsilon)}\left(\frac{\tilde{W}_n(\tau)}{\tilde{W}(\varepsilon)} - 1\right).$$

Along with $\mathbb{P}_0(T_n^{-\varepsilon} < T_n^{(0,\infty)}) = \frac{\tilde{W}_n(0)}{\tilde{W}_n(\varepsilon)}$ and $\mathbb{P}_{-\varepsilon}(T_n^{(0,\infty)} < T_n^{-\tau}) = 1 - \frac{\tilde{W}_n(\varepsilon)}{\tilde{W}_n(\tau)}$, this gives

$$\begin{aligned}
&\mathbb{P}_0(\Upsilon_n - \varepsilon \in \mathrm{d}u,\ T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}) \\
=&\frac{\tilde{W}_n(\tau-\varepsilon)}{\tilde{W}_n(\tau)}\ \int_{z\in(u,\infty)} N_n'(-\epsilon(\chi-) \in \mathrm{d}u,\ \epsilon(\chi) - \epsilon(\chi-) \in \mathrm{d}z,\ -\inf\epsilon < \tau - \varepsilon,\ \sup\epsilon \geq \varepsilon) \\
=&\frac{\tilde{W}_n(\tau-\varepsilon)}{\tilde{W}_n(\tau)}\ \frac{n}{d_n}e^{-\tilde{\eta}_n u}\mathrm{d}u\ \mathbb{P}_u(T_n^0 < T_n^{(\tau-\varepsilon,\infty)} \mid T_n^0 < \infty)\ \int_{z\in(u,\infty)} \tilde{\Lambda}_n(\mathrm{d}z)\mathbb{P}_{z-u}(T_n^{(\varepsilon,\infty)} < T_n^0),
\end{aligned}$$

where in the last equality we first appealed to the strong Markov property at $T_n^{(0,\infty)}$, and then to Proposition 0.7. Finally, we get

$$\begin{aligned}
&\mathbb{P}_0(\Upsilon_n - \varepsilon \in \mathrm{d}u,\ T_n^{-\varepsilon} < T_n^{(0,\infty)} < T_n^{-\tau}) \\
&\qquad = \frac{n}{d_n}\ \frac{\tilde{W}_n(\tau-u-\varepsilon)}{\tilde{W}_n(\tau)}\ \mathrm{d}u\ \int_{z\in(u,\infty)} \tilde{\Lambda}_n(\mathrm{d}z)\left(1 - \frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right),
\end{aligned}$$

which, along with ${p'}_{n,\varepsilon} = \frac{d_n}{n}\left(\frac{1}{\tilde{W}_n(\varepsilon)} - \frac{1}{\tilde{W}_n(\tau)}\right)$, proves (16).

We next want to prove (17). A similar reasoning as for (16) holds, except that Z has infinite variation. In particular, if Z has a Gaussian component, the process can then creep upwards. Using as before the strong Markov property at $T^{-\varepsilon}$ and Proposition 0.5.2(ii) in [Ber96], we have

$$\begin{aligned}&N'(\Upsilon-\varepsilon\in \mathrm{d}u,\ -\inf\epsilon\in[\varepsilon,\tau))\\&\quad= N'(-\inf\epsilon>\varepsilon)\mathbb{P}_{-\varepsilon}(T^{(0,\infty)}<T^{-\tau})\frac{N'(-\epsilon(\chi-)\in \mathrm{d}u,\ -\inf\epsilon<\tau-\varepsilon,\ \sup\epsilon\geq\varepsilon)}{N'(-\inf\epsilon<\tau-\varepsilon,\ \sup\epsilon\geq\varepsilon)}.\end{aligned}$$

On the one hand, from [OP09, Section 4] we know that

$$\begin{aligned}N'(-\inf\epsilon>\varepsilon)&=\frac{1}{W(\varepsilon)},\\N'(-\inf\epsilon<\tau-\varepsilon,\ \sup\epsilon\geq\varepsilon)&=\frac{1}{W(\tau-\varepsilon)}\left(\frac{W(\tau)}{W(\varepsilon)}-1\right),\end{aligned}$$

which gives

$$\mathbb{P}_{-\varepsilon}(T^{(0,\infty)}<T^{-\tau})\frac{N'(-\inf\epsilon>\varepsilon)}{N'(-\inf\epsilon<\tau-\varepsilon,\ \sup\epsilon\geq\varepsilon)}=\left(1-\frac{W(\varepsilon)}{W(\tau)}\right)\frac{W(\tau-\varepsilon)}{W(\tau)-W(\varepsilon)}=\frac{W(\tau-\varepsilon)}{W(\tau)}.$$

On the other hand, distinguishing the excursions entering $(0,\infty)$ immediately from the others leads to

$$\begin{aligned}&N'(-\epsilon(\chi-)\in \mathrm{d}u,\ -\inf\epsilon<\tau-\varepsilon,\ \sup\epsilon\geq\varepsilon)\\&\quad= N'(-\epsilon(\chi-)\in \mathrm{d}u,\ -\inf\epsilon\in(0,\tau-\varepsilon),\ \sup\epsilon\geq\varepsilon)+N'(-\inf\epsilon=0,\ \sup\epsilon\geq\varepsilon)\delta_0(\mathrm{d}u),\end{aligned}$$

where $N'(-\inf\epsilon=0,\ \sup\epsilon\geq\varepsilon)=\frac{b^2}{2}\frac{W'(\varepsilon)}{W(\varepsilon)}$ according to [OP09, Section 4].
Then similarly as before, applying the strong Markov property at $T^{(0,\infty)}$ and Proposition 0.7, we get

$$\begin{aligned}&N'(\Upsilon-\varepsilon\in \mathrm{d}u,\ -\inf\epsilon\in[\varepsilon,\tau))\\&=\frac{W(\tau-\varepsilon)}{W(\tau)}\Bigg(e^{-\tilde{\eta}u}\mathrm{d}u\ \mathbb{P}_u(T^0<T^{(\tau-\varepsilon,\infty)}\,|\,T^0<\infty)\int_{z\in(u,\infty)}\Lambda(\mathrm{d}z)\mathbb{P}_{z-u}(T^{(\varepsilon,\infty)}<T^0)\\&\qquad\qquad+\frac{b^2}{2}\frac{W'(\varepsilon)}{W(\varepsilon)}\delta_0(\mathrm{d}u)\Bigg)\\&=\frac{W(\tau-\varepsilon)}{W(\tau)}\left(\frac{b^2}{2}\frac{W'(\varepsilon)}{W(\varepsilon)}\delta_0(\mathrm{d}u)+\mathrm{d}u\ \frac{W(\tau-u)}{W(\tau-\varepsilon)}\int_{z\in(u,\infty)}\Lambda(\mathrm{d}z)\left(1-\frac{W(\varepsilon-(z-u))}{W(\varepsilon)}\right)\right),\end{aligned}$$

which proves (17). □

Proof of Lemma 3.24 :
We can now prove the weak convergence of $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ to $\nu_{\varepsilon}^{\mathrm{INIT}}$. To begin with, formulas (16) and (17) of Lemma 3.23, along with the convergence of $\tilde{W}_n$ towards W (which implies in particular ${p'}_{n,\varepsilon}\to p_\varepsilon$ as $n\to\infty$), ensure that we only have to prove the weak convergence of

$$\mathrm{d}u\ \mathbb{1}_{u\in(0,\tau-\varepsilon)}\ \tilde{W}_n(\tau-u-\varepsilon)\int_{z\in(u,\infty)}\tilde{\Lambda}_n(\mathrm{d}z)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right)$$

towards

$$\delta_0(\mathrm{d}u)\frac{b^2}{2}\frac{W'(\varepsilon)}{W(\varepsilon)}W(\tau-\varepsilon)+\mathrm{d}u\ \mathbb{1}_{u\in(0,\tau-\varepsilon)}\ W(\tau-u-\varepsilon)\ \int_{z\in(u,\infty)}\Lambda(\mathrm{d}z)\left(1-\frac{W(\varepsilon-(z-u))}{W(\varepsilon)}\right).$$

First notice that since $\tilde{W}_n$ and W vanish on the negative half-line, we have

$$\int_{(u,\infty)}\tilde{\Lambda}_n(\mathrm{d}z)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right)=\bar{\tilde{\Lambda}}_n(u+\varepsilon)+\int_{(u,u+\varepsilon]}\tilde{\Lambda}_n(\mathrm{d}z)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right)$$

$$\text{and }\int_{(u,\infty)}\Lambda(\mathrm{d}z)\left(1-\frac{W(\varepsilon-(z-u))}{W(\varepsilon)}\right)=\bar{\Lambda}(u+\varepsilon)+\int_{(u,u+\varepsilon]}\Lambda(\mathrm{d}z)\left(1-\frac{W(\varepsilon-(z-u))}{W(\varepsilon)}\right).$$

The functions $u\mapsto\tilde{W}_n(\tau-u-\varepsilon)\bar{\tilde{\Lambda}}_n(u+\varepsilon)$ converge pointwise on $(0,\tau-\varepsilon)$ towards $u\mapsto W(\tau-u-\varepsilon)\bar{\Lambda}(u+\varepsilon)$ and are bounded by $\sup_{n\geq1}\bar{\tilde{\Lambda}}_n(\varepsilon)$. Then by dominated convergence, we have the following weak convergence :

$$\mathrm{d}u\ \tilde{W}_n(\tau-u-\varepsilon)\bar{\tilde{\Lambda}}_n(u+\varepsilon)\mathbb{1}_{u\in(0,\tau-\varepsilon)}\ \Rightarrow\ \mathrm{d}u\ W(\tau-u-\varepsilon)\bar{\Lambda}(u+\varepsilon)\mathbb{1}_{u\in(0,\tau-\varepsilon)}.$$

Finally, it remains to prove the weak convergence of

$$\mathrm{d}u\mathbb{1}_{u\in(0,\tau-\varepsilon)}\tilde{W}_n(\tau-u-\varepsilon)\int_{(u,u+\varepsilon]}\tilde{\Lambda}_n(\mathrm{d}z)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right)$$

towards

$$\delta_0(\mathrm{d}u)\frac{b^2}{2}\frac{W'(\varepsilon)}{W(\varepsilon)}W(\tau-\varepsilon)+\mathrm{d}u\ \mathbb{1}_{u\in(0,\tau-\varepsilon)}\ W(\tau-u-\varepsilon)\ \int_{z\in(u,u+\varepsilon]}\Lambda(\mathrm{d}z)\left(1-\frac{W(\varepsilon-(z-u))}{W(\varepsilon)}\right).$$

Consider g a continuous bounded function on $\mathbb{R}_+$. We have :

$$\begin{aligned}
&\int_0^{\tau-\varepsilon}\mathrm{d}u\ g(u)\ \tilde{W}_n(\tau-u-\varepsilon)\left(\int_{(u,u+\varepsilon]}\tilde{\Lambda}_n(\mathrm{d}z)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right)\right)\\
&=\int_{(0,\tau]}\tilde{\Lambda}_n(\mathrm{d}z)\left(\int_{0\wedge z-\varepsilon}^{z}\mathrm{d}u\ g(u)\tilde{W}_n(\tau-u-\varepsilon)\left(1-\frac{\tilde{W}_n(\varepsilon-(z-u))}{\tilde{W}_n(\varepsilon)}\right)\right)\\
&=\int_{(0,\tau]}\tilde{\Lambda}_n(\mathrm{d}z)\left(\int_0^{z\wedge\varepsilon}\mathrm{d}v\ g(z-v)\tilde{W}_n(\tau-\varepsilon-z+v)\left(1-\frac{\tilde{W}_n(\varepsilon-v)}{\tilde{W}_n(\varepsilon)}\right)\right).
\end{aligned}$$

We set, for all $z\geq0$,

$$\begin{aligned}
h_n(z)&:=\textstyle\int_0^{z\wedge\varepsilon}\mathrm{d}v\ g(z-v)\tilde{W}_n(\tau-\varepsilon-z+v)\big(1-\frac{\tilde{W}_n(\varepsilon-v)}{\tilde{W}_n(\varepsilon)}\big)\\
h(z)&:=\textstyle\int_0^{z\wedge\varepsilon}\mathrm{d}v\ g(z-v)W(\tau-\varepsilon-z+v)\big(1-\frac{W(\varepsilon-v)}{W(\varepsilon)}\big).
\end{aligned}$$

We then verify that the conditions of Proposition I.4.4, are fulfilled :

- The functions h_n and h can be bounded by $\varepsilon\cdot\sup|g|$ and are continuous thanks to the continuity on $\mathbb{R}_+$ of the functions $\tilde{W}_n$ and W.

- The dominated convergence theorem and the uniform convergence of $\tilde{W}_n$ towards W on $\mathbb{R}_+$ (see Proposition 3.1.(iii)) ensure that h_n converges uniformly on $\mathbb{R}_+$ towards h (recall that $W(\varepsilon)>0$).

- Now since $\tilde{W}_n'$ converges uniformly towards W' on every compact set of $\mathbb{R}_+^*$ (again from Proposition 3.1.(iii)), the sequence $(v \mapsto \frac{1}{v}(\tilde{W}_n(\varepsilon) - \tilde{W}_n(\varepsilon - v)))_n$ converges uniformly towards $v \mapsto \frac{1}{v}(W(\varepsilon) - W(\varepsilon - v))$ on $(0, \frac{\varepsilon}{2})$. Consequently, for all $a > 0$, if n is large enough we have

$$\sup_{u\in(0,\varepsilon/2)} \left| \frac{h_n(u) - h(u)}{u^2} \right| \leq \sup_{u\in(0,\varepsilon/2)} \frac{a}{u^2} \sup|g| \int_0^u v \, \mathrm{d}v = \frac{a}{2} \sup|g|,$$

and thus we have uniform convergence of $u \mapsto h_n(u)/u^2$ towards $u \mapsto h(u)/u^2$ on $(0, \varepsilon/2)$.

- In the same way we get from the continuity of g, $\tilde{W}_n$ and W that

$$\frac{h(u)}{u^2} \underset{u\to 0}{\to} \frac{1}{2} g(\varepsilon) \frac{W'(\varepsilon)}{W(\varepsilon)} W(\tau - \varepsilon).$$

We then get the expected convergence from an appeal to Proposition I.4.4. As a conclusion, we proved that the measures $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ converge weakly to $\nu_\varepsilon^{\mathrm{INIT}}$. □

Lemma 3.25. *As $n \to \infty$, $M_{n,\varepsilon}(0)$ converges in distribution towards $M_\varepsilon(0)$.*

Proof :
We have to prove the weak convergence of

$$\nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}u \times \{1\}), \quad \int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x \times \{0\}) \nu_n^{\mathrm{M}}(x, \mathrm{d}u) \quad \text{and} \quad \int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x, \{0\}) \nu_n^{\mathrm{D}}(x, \mathrm{d}u).$$

The convergence of the first one is a straightforward consequence of Lemma 3.24. We prove below the convergence of the Laplace transform of the second one, and a similar reasoning holds for the third one. We consider for all $a \geq 0$

$$\int_{(0,\infty)} e^{-au} \int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x \times \{0\}) \nu_n^{\mathrm{M}}(x, \mathrm{d}u) = \int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x \times \{0\}) \int_{[\varepsilon,\tau)} e^{-au} \nu_n^{\mathrm{M}}(x, \mathrm{d}u)$$

and set $h_n(x) := \int_{[\varepsilon,\tau)} e^{-au} \nu_n^{\mathrm{M}}(x, \mathrm{d}u)$ and $h(x) := \int_{[\varepsilon,\tau)} e^{-au} \nu^{\mathrm{M}}(x, \mathrm{d}u)$.

The functions h and h_n are all bounded by 1. Moreover, they are continuous : indeed, we have

$$|\mathbb{E}(e^{-aH^+(\mathrm{e})},\ L^{-1}(\mathrm{e}) < T^{-x}) - \mathbb{E}(e^{-aH^+(\mathrm{e})},\ L^{-1}(\mathrm{e}) < T^{-x_0})| \leq \mathbb{P}(T^{-x_0} < T^{-x}),$$

which vanishes as $x \to x_0$ thanks to the a.s. continuity of $x \mapsto T^{-x}$ on $\mathbb{R}_+$ under $\mathbb{P}$. Here again, for the sake of simplicity, we omitted the conditioning, but a similar reasoning and an appeal to the continuity of W, lead to the continuity of h. Besides, the same arguments can be used to get the continuity of h_n. Finally, as established in the proof of [Kar75, Th.4], Lemma 3.22 ensures the uniform convergence of h_n towards h on every compact set of $\mathbb{R}_+^*$. Then, since $\nu_{n,\varepsilon}^{\mathrm{INIT}}$ and $\nu_\varepsilon^{\mathrm{INIT}}$ are probability measures such that $\nu_{n,\varepsilon}^{\mathrm{INIT}} \Rightarrow \nu_\varepsilon^{\mathrm{INIT}}$, Lemma A.1 entails the convergence of the Laplace transform of $\int_{[\varepsilon,\tau)} \nu_{n,\varepsilon}^{\mathrm{INIT}}(\mathrm{d}x \times \{0\}) \nu_n^{\mathrm{M}}(x, \mathrm{d}u)$ towards that of $\int_{[\varepsilon,\tau)} \nu_\varepsilon^{\mathrm{INIT}}(\mathrm{d}x \times \{0\}) \nu^{\mathrm{M}}(x, \mathrm{d}u)$. □

Proof of Proposition 3.9 :
Lemma 3.22 ensures that the Markov chains $M_{n,\varepsilon}$ and M_ε satisfy condition (4).b in [Kar75], while condition (4).a of the same paper is given by Lemma 3.25. Then the announced convergence is a consequence of [Kar75, Theorem (1)]. □

A Appendix

A.1 A convergence lemma for integrals

Lemma A.1. *Let $(h_n)_{n\geq 0}$ and h be continuous bounded mappings from $\mathbb{R}^d$ to $\mathbb{R}$, and suppose (h_n) is dominated by a bounded function. Let $(\mu_n)_{n\geq 0}$ and μ be in $\mathcal{M}_f(\mathbb{R}^d)$ and suppose that*

(i) *(μ_n) converges weakly to μ.*

(ii) *The sequence of mappings (h_n) converges to h uniformly on every compact set of $\mathbb{R}^d$.*

Then

$$\int h_n \mathrm{d}\mu_n \xrightarrow{n\to\infty} \int h \mathrm{d}\mu$$

Proof :
We have

$$\left|\int h_n \mathrm{d}\mu_n - \int h \mathrm{d}\mu\right| \leq \left|\int (h_n - h)\mathrm{d}(\mu_n - \mu)\right| + \left|\int (h_n - h)\mathrm{d}\mu\right| + \left|\int h \mathrm{d}(\mu_n - \mu)\right|.$$

The mapping h is continuous and bounded on $\mathbb{R}^d$, then (i) implies the convergence to 0 of the term $|\int h\mathrm{d}(\mu_n - \mu)|$. The domination and convergence assumptions made on (h_n) allow us to apply the dominated convergence theorem to get the convergence of $|\int (h_n - h)\mathrm{d}\mu|$ to 0. As for the first term in the sum, it requires some additional details : Let ε be a positive real number. First, thanks to (i) and since $(h_n - h)$ is dominated by a constant, we can find a compact set $K_\varepsilon \subset \mathbb{R}^d$ and $n_0 \in \mathbb{N}$ such that $|\int_{K_\varepsilon^c} (h_n - h)\mathrm{d}(\mu_n - \mu)| \leq \varepsilon$ for $n \geq n_0$. Secondly the uniform convergence on the compact set K_ε of the sequence (h_n) ensures that $|\int_{K_\varepsilon} (h_n - h)\mathrm{d}(\mu_n - \mu)| \leq \varepsilon$ for n large enough. In consequence we have convergence of the term $|\int (h_n - h)\mathrm{d}(\mu_n - \mu)|$ to 0, and the result follows. □

A.2 Consequences of Assumption A

Proof of Proposition 3.1 :

(i) Since Z is a.s. continuous at T^{-x} (resp. is not a compound Poisson process), we have $\lim_{\varepsilon\to 0+} T^{-(x+\varepsilon)} = T^{-x}$ (resp. $\lim_{\varepsilon\to 0+} T^{(y+\varepsilon,\infty)} = T^{(y,\infty)}$) a.s., and hence the convergence in law of T_n^{-x} towards T^{-x} (resp. $T_n^{(y,\infty)}$ towards $T^{(y,\infty)}$) is a straightforward consequence of Proposition VI.2.11 in [JS87].

(ii) Now ϕ_n (resp. ϕ) is the Laplace exponent of the process $x \mapsto T_n^{-x}$ (resp. $x \mapsto T^{-x}$) [Ber96, Th.VII.1.1]. The pointwise convergence of $\tilde{\phi}_n$ to ϕ is thus a consequence of point (i). The uniform convergence comes from the fact that for all $n \geq 1$, $\tilde{\phi}_n$ is increasing on $\mathbb{R}_+$.

(iii) The proof of the pointwise convergence of $\tilde{W}_n$ (resp. $\tilde{W}_n'$) towards W (resp. W') can be found in [LS12, Prop.3.1]. Moreover, we have for all $y > x$ $\mathbb{P}(T^{-x} < T^{(y-x,\infty)}) = \frac{W(x)}{W(y)}$ [Ber96, Th.VII.2.8], and then the function $x \mapsto \tilde{W}_n(x)/\tilde{W}_n(y)$ is decreasing, thus the convergence of $\tilde{W}_n$ towards W is uniform on every compact set of $\mathbb{R}_+$. Finally, the uniform convergence of $\tilde{W}_n'$ towards W' on every compact set of $\mathbb{R}_+^*$ can be deduced from the expression of $\tilde{W}_n'$ given in the proof of Lemma 8.2 in [Kyp06], as a product of two monotone functions. □

Chapter III

Sample genealogy and mutational patterns for critical branching populations

The article [ADL14] is joint work with Guillaume Achaz and Amaury Lambert.

1 Introduction

A major concern in population genetics is the prediction of patterns of genetic variation with help of stochastic models. The reference model currently used by biologists to answer this question is the Kingman coalescent model [Kin82b, Kin82a] coupled with Poissonian mutations on the lineages. As the scaling limit of numerous constant population size models, such as Wright-Fisher and Moran models, it encompasses the two population models that are most commonly used by biologists. The genealogical structure of a sample (rather than of the total population) is well-known (equivalently given by the Kingman coalescent), and explicit results on the allelic partition generated by rare, neutral mutations (equivalent to a Kingman coalescent with Poissonian mutations) are provided by Ewens' sampling formula [Ewe72, Dur08]. In this work, we intend to study the genealogical and mutational patterns of a sample from a branching population, in order to offer an alternative model where the constant population size assumption is released, with no *a priori* assumption on the variation of the population size over time. The sampling is here essential to make the model applicable to real data and comparable to the Kingman coalescent model.

The genealogy of branching populations was in particular studied by L. Popovic in [Pop04], in the setting of the critical birth-death process conditioned on its population size at a fixed time horizon, and later by A. Lambert in [Lam10] in the more general framework of splitting trees. The genealogy of the extant individuals is described as a random point process, called *coalescent point process*, which distribution is characterized by a sequence of i.i.d. random variables.

Here we want to focus on the genealogy of a sample rather than of the total extant population. The question of sampling in birth-death models has already been approached with two

different points of view. On the one hand, [Pop04] and [Sta09] deal with Bernoulli sampling of the total population. This approach rather applies to the species scale, for example in the case of incomplete phylogenies. On the other hand, in [Sta08] and [Sta09], T. Stadler considers the case of a uniform sample of m individuals among the extant ones, in the birth-death process conditioned on its population size at present time, with uniform prior on its time of origin. Our approach is based on Bernoulli sampling with conditioning on the sample size, in order to get a uniform sample with fixed size without having to condition on the total extant population size.

We first consider in Section 1.1 sample genealogies in a general framework of branching populations with neutral mutations at birth. We make use of convergence results obtained by one of the authors [Del13b] (Chapter II in this thesis) to show how a broad class of such populations all result in the same distribution for the genealogy of a sample, namely the law of a critical birth-death model with Poissonian mutations on the lineages. We then specify in Section 1.2 the model that we adopt for the rest of the paper. We finally present in Section 1.3 the outline and the main results of this work : in Section 1.3.1, we investigate the law of the genealogy of a sample in the critical birth-death model conditioned on its sample size, with various prior distributions on the foundation time of the population. We provide in Section 1.3.2 explicit formulae for the expected site frequency spectrum of the sample. Section 1.3.3 is then devoted to the convergence in distribution of the sample genealogy, as the sample size gets large. Furthermore, we state that the limiting genealogies with different priors can all be embedded in the same realization of a given Poisson point measure.

1.1 Genealogies and sampling in branching populations conditioned on survival

Let us first consider a very general model of branching populations with mutations : let $(\mathbb{T}_N)_{N\in\mathbb{N}}$ be a sequence of splitting trees, i.e. random trees where individuals have lifetimes that do not necessarily follow an exponential distribution, during which they give birth at constant rate to i.i.d copies of themselves [Gei96, GK97, Lam10]. For any N, $\mathbb{T}_N$ is characterized by its so-called *lifespan measure* Λ_N, which is a σ-finite measure on $(0,\infty)$ such that $\int(1\wedge r)\Lambda_N(\mathrm{d}r) < \infty$. We further assume that any individual in $\mathbb{T}_N$ experiences, conditional on her lifetime r, a mutation at birth with probability $f_N(r)$, where f_N is a continuous function from $\mathbb{R}_+^*$ to $[0,1]$ called *mutation function*. We adopt the classical assumptions of neutral mutations (i.e. mutations do not affect the population dynamics) and of the infinite-site model [Kim69] : each individual is associated to a DNA sequence, and each mutation occurs at a site that has never mutated before.

Finally, we fix $t > 0$, and we condition $\mathbb{T}_N$ on survival at time Nt. We work later in a time scale where a unit of time is proportional to N : the factor N can thus be seen as a counterpart of the constant population size of the Wright-Fisher model. We assume that each individual alive at Nt is independently sampled with probability $p_N \in (0,1)$. Individuals are labeled according to the order defined in Section 0.1.4 (« left to right » order associated to the planar representation of the tree when daughters all sprout to the right of their mother), and we denote by $I_N = (I_{Nj})_j$ the sequence of indices of the sampled individuals. See Figure 1 for a graphical representation of $\mathbb{T}_N$, and of some objects hereafter defined.

We are here interested in the distribution of the genealogy of the sampled individuals in $\mathbb{T}_N$, and we consider the model under two slightly different points of view : in case (I), relying on results of Chapter II, we consider a scaling limit in a large population asymptotic, while in case (II), we consider the example of the critical birth-death process, for which results can be obtained without necessarily having to consider $N \to \infty$. We show here how these two settings

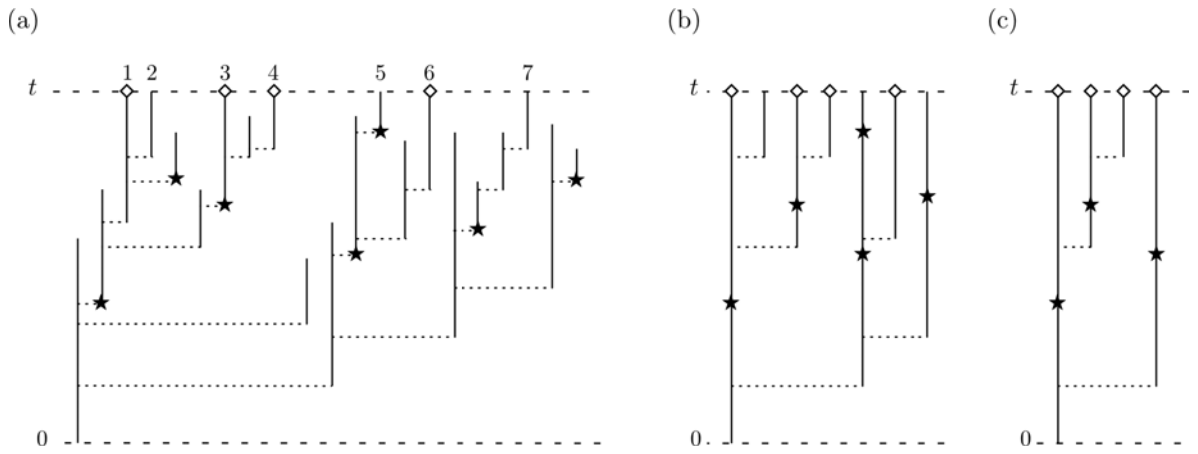

FIGURE 1 – In the three panels (a), (b), (c), the vertical axis indicates time. The horizontal (dotted) lines show filiation. Mutations are symbolized by $\star$ and sampled individuals by $\diamond$.
(a) An example of the rescaled tree $\mathbb{T}_N$ with 7 extant individuals at t, where 4 individuals are sampled.
(b) its (marked) coalescent point process (later referred to as Σ_N),
(c) and the (marked) coalescent point process of the sampled individuals.

lead to the same distribution for the genealogy of a sample, justifying hence the model we later consider for the rest of the paper.

To this aim we rescale time in $\mathbb{T}_N$ by multiplying all the edge lengths of $\mathbb{T}_N$ by a factor $1/N$. This rescaled tree is still denoted by $\mathbb{T}_N$, and is now originating at time t. Then we introduce, for any $N \in \mathbb{N}$, the so called *marked coalescent point process* Σ_N (see Chapter II), i.e. the tree spanned by the genealogy of the extant population of $\mathbb{T}_N$ at time t, enriched with the mutational history of extant individuals. More precisely, Σ_N is a point measure that can be expressed as $\Sigma_N = \sum_{i=1}^{\mathcal{N}-1} \delta_{(i,\sigma_i^N)}$ where $\mathcal{N}$ is the number of extant individuals at time t, and for any $1 \leq i \leq \mathcal{N}-1$, σ_i^N is itself a point measure, whose set of atoms contains, in addition to the coalescence time between individuals i and $i+1$, all the times at which a mutation occurred on the i-th lineage (see Figure 1).

(I) Scaling limit.
First, we assume that $(\mathbb{T}_N)$ converges, as $N \to \infty$, towards a Brownian tree (see e.g. [Ald93]) : for any $N \in \mathbb{N}$, for any $\lambda \geq 0$, define $\psi_N(\lambda) := -\lambda - \int_{(0,\infty)}(1 - e^{-\lambda r})\Lambda_N(\mathrm{d}r)$. We assume that the sequence $(\mathbb{T}_N)$ follows (a particular case of) Assumption A of Chapter II :

> *Assumption A : There exists a sequence of positive real numbers $(d_N)_{N\geq 1}$ such that as $N \to \infty$, the sequence $(d_N \psi_N(\cdot/N))$ converges towards $\lambda \mapsto \lambda^2$, i.e. the Laplace exponent of a Brownian motion.*

This assumption has to be interpreted as the convergence in law of the so-called *jumping chronological contour process* of the rescaled tree $\mathbb{T}_N$, which distribution is characterized by a Lévy process with finite variation, drift -1 and Lévy measure Λ_N (see Section 0.1.4).
Second, we fix $\theta \in \mathbb{R}_+$ and we suppose that the sequence of mutation functions (f_N) satisfies one of the following convergence assumptions (see Chapter II) :

Assumption B.1 : For all $N \geq 1$, for all $u \in \mathbb{R}_+$, $f_N(u) = \theta_N$, where $\theta_N \in [0,1]$ is such that $\frac{d_N}{N}\theta_N \underset{N\to\infty}{\longrightarrow} \theta$.

Assumption B.2 : The sequence $\left(u \mapsto \frac{f_N(Nu)}{1\wedge u}\right)$ converges uniformly to $u \mapsto \frac{f(u)}{1\wedge u}$ on $\mathbb{R}_+^*$, where f is a continuous function from $\mathbb{R}_+$ to $\mathbb{R}_+$ satisfying $f(u)/u \to \theta$ as $u \to 0^+$.

Then we have the following convergence.

Theorem. *[Del13b, Th.3.2] The (space rescaled) point measure $\Sigma_N = \sum_{i=1}^{\mathcal{N}-1} \delta_{(i\frac{N}{d_N}, \sigma_i^N)}$ converges in distribution, as $N \to \infty$, towards a Poisson point process on $[0,\mathrm{e}] \times (0,t)$ with intensity $\mathrm{d}l\, x^{-2}\mathrm{d}x$, where e is an independent exponential variable with parameter $1/t$, with independent Poissonian mutations at rate θ on the lineages.*

Besides, we assume that the sampling probability is given by $p_N = p\,N/d_N$, where p is a fixed positive real number such that p_N is in $(0,1)$ for N large enough. Then the rescaled sequence $(\frac{N}{d_N} I_N)$ of indices of the sampled individuals (independent of $(\mathbb{T}_N)$), converges towards the sequence of jump times of an independent Poisson process with rate p. The joint convergence of Σ_N with $\frac{N}{d_N} I_N$ is of course provided by their independence.

As a consequence, from [Lam08] we deduce that the coalescent point process of the sampled individuals is then distributed as the coalescent point process of a critical birth-death model with rate p conditioned on survival at time t, with independent Poissonian mutations at rate θ on the lineages.

(II) Critical birth-death tree.
Second, fix $N \in \mathbb{N}$, $p \in (0,N)$, and consider the example where $\mathbb{T}_N$ is a critical birth-death tree with rate N conditioned on survival at time t. Then, set $p_N = p/N$ and assume that the mutation function f_N is constant, equal to θ/N. This is in fact a particular case of (I) (Assumptions A and B.1 are satisfied with $d_N = N^2$), but here we do not need to let $N \to \infty$. For any $N \in \mathbb{N}$, the marked coalescent point process Σ_N is distributed as the coalescent point process of a critical birth-death model with rate 1 conditioned on survival at time t, with Poissonian mutations at rate θ on the lineages (see [Pop04, Sec.3] and [Del13b, Ex.1]). Finally, from [Lam08], we get that the coalescent point process of the sample is then distributed, exactly as above, as the coalescent point process of a critical birth-death model with rate p conditioned on survival at time t, with independent Poissonian mutations at rate *theta* on the lineages.

Since the two cases (I) and (II) result in the same distribution for the genealogy of a sample, we limit our study to case (II). Besides, since the mutation schemes arise as independent of genealogies, the results concerning distributions of genealogies are stated without reference to mutations.

1.2 Model with conditioning on the sample size

From now on, consider $\mathbb{T}$ a critical birth-death tree with rate 1. Time is now counted backwards into the past, i.e. « present time » is now time 0, and « u units of time before present » is now time u. We begin with the case of a fixed foundation time of the population. The model has four parameters : a time $t \in \mathbb{R}_+^*$, a scaling factor $N \in \mathbb{R}_+^*$, a positive integer n (the sample size), and a sampling parameter $p \in (0,N)$.

Assume first that $\mathbb{T}$ has been founded Nt units of time ago. As previously, individuals are independently sampled at present time, with probability p/N. Besides, we rescale time by a factor $1/N$ (all the edge lengths are then multiplied by a factor $1/N$). We keep the notation $\mathbb{T}$ for the rescaled tree, so that $\mathbb{T}$ is now a critical birth-death tree with rate N, originating at time t.

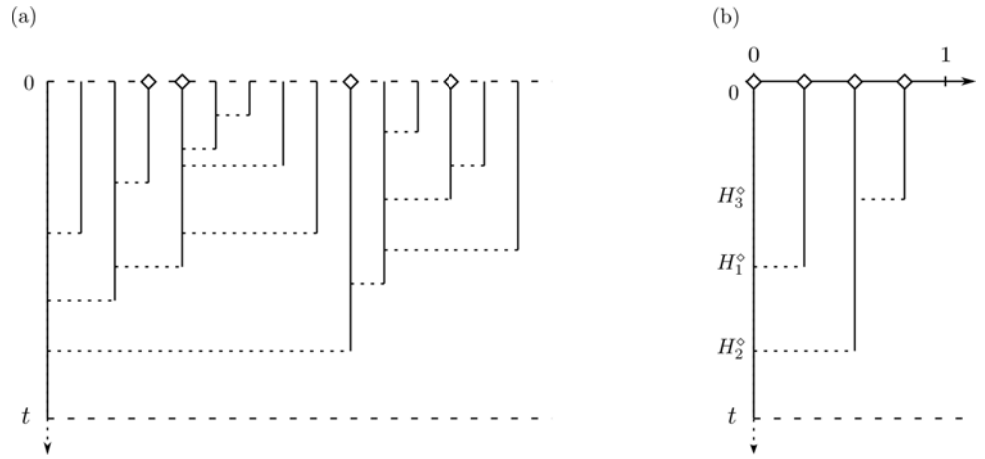

FIGURE 2 – In both figures (a) and (b), the vertical axis indicates time (running backwards).
(a) A graphical representation of the coalescent point process at present time of a (rescaled) tree $\mathbb{T}$ originating at time t with 15 extant individuals and $n = 4$ sampled individuals (symbolized by $\diamond$). The horizontal lines show filiation.
(b) A graphical representation of the coalescent point process $\pi_n = \sum_{k=1}^{n-1} \delta_{\left(\frac{k}{n}, H_k^\diamond\right)}$ of the sample represented in (a).

We now introduce the conditioning on the sample size : we condition $\mathbb{T}$ on having n sampled individuals at present time. Note that after conditioning, the distribution of the n sampled individuals within the total extant population does not depend on p, and is a posteriori equivalent to uniform, sequential sampling.

The genealogy of the n sampled individuals is characterized by its coalescent point process

$$\pi_n = \sum_{k=1}^{n-1} \delta_{\left(\frac{k}{n}, H_k^\diamond\right)},$$

where for $1 \leq k \leq n-1$, $H_k^\diamond$ is the divergence time between the k-th and the $(k+1)$-th sampled individual in the rescaled tree $\mathbb{T}$ (see Figure 2). The space rescaling by a factor $1/n$ ensures in particular that the supports of the measures π_n converge as $n \to \infty$, which is required by the results later established in the large sample size asymptotic. Besides, recall that thanks to their independence with the genealogy, mutations are for now deliberately omitted. Finally, we define $(T_{n,k})_{1 \leq k \leq n-1}$ the decreasing reordering of the divergence times $(H_k^\diamond)_{1 \leq k \leq n-1}$.

1.3 Outline and statement of results

The first purpose of this paper is to study the distribution of π_n, under various hypotheses on the origin of the process : we denote by

- $\mathbb{P}_n^t$ the law of the rescaled tree $\mathbb{T}$ with fixed time of origin t and sample size n,

- $\mathbb{P}_n^{(\infty)}$ the law of $\mathbb{T}$ with infinite time of origin and sample size n,

- $\mathbb{P}_n^{(i)}$ the law of $\mathbb{T}$ with random time of origin, with (potentially improper) prior distribution $g_i : x \mapsto x^{-i}$, $i \in \mathbb{Z}_+$, and sample size n.

Note that the case $i = 0$ corresponds to the case of a uniform prior investigated in [AP05] and [Ger08]. This study, presented in Section 2, will then enable us to derive results concerning mutational patterns of the sample (Section 3), and then concerning the behaviour of the genealogy as the sample size gets large (Section 4).

1.3.1 A universal law for the genealogy of a sample

First, in the case of a fixed time of origin, the law of π_n under $\mathbb{P}_n^t$ is independent of N and is specified by the following result (Theorem 2.1) :

Theorem. *Under $\mathbb{P}_n^t$, $(H_i^\diamond)_{1\leq i\leq n-1}$ is a sequence of i.i.d. random variables with probability density function $x \mapsto \frac{p}{(1+px)^2}\frac{1+pt}{pt}\mathbb{1}_{(0,t)}(x)$. In other words, the coalescent point process π_n has the law of the genealogy of a critical birth-death tree with rate p conditioned on having n extant individuals at time t.*

We then prove that this equality in law still holds when letting the time go to infinity or when randomizing the time of origin (with prior distribution g_i, $i \in \mathbb{Z}_+$) in both processes : for example, under $\mathbb{P}_n^{(i)}$ the coalescent point process π_n has the law of the genealogy of a critical birth-death tree with rate p, with prior g_i on its time of origin, and conditioned on having n extant individuals at present time. Hence whatever the assumption on the foundation time of the population, the study of the genealogy of the sample boils down to the same object : the genealogy of a critical birth-death process with rate p, with extant population size n.
Following on from results provided by [Ger08] in the case of a uniform prior, we then obtain the following property for the successive divergence times $(T_{n,k})_{1\leq k\leq n-1}$ (Proposition 2.9) :

Proposition. *Under $\mathbb{P}_n^{(i)}$, the time $T_{n,k}$ to the k-th most recent common ancestor has finite moment of order m iff $m \leq k + i$.*

Although we limited here our study to the framework (II) introduced earlier, one could certainly generalize these results (and the upcoming ones) to the scaling limit of case (I). To prove this, one would have to consider a sequence of trees conditioned on their sample size, and then to establish the convergence, in the large population asymptotic, of the marked coalescent point process of the sample. This is however beyond the scope of the present paper.

1.3.2 Mutational patterns

In Section 3, we study the so-called *site frequency spectrum* of the sample, i.e. the $(n-1)$-tuple $(\xi_1, \ldots, \xi_{n-1})$, where ξ_k is the number of mutations carried by k individuals in the sample. Various results for the frequency spectrum in the framework of general branching processes are established in [Lam08, CL12a, CL12b, Ric14]. One of the authors investigates in [Lam08] the case of coalescent point processes with Poissonian mutations on germ lines, and obtains asymptotic results for the site and allele frequency spectrum of large samples. Explicit formulae for the expected allele frequency spectrum of a splitting tree with n individuals at fixed time horizon t are provided by N. Champagnat and this author in the case of Poissonian mutations on the lineages [CL12a], and by M. Richard in the case of mutations at birth [Ric14]. Their results are

compared in [CLR12] in the particular case of birth-death processes. Further results about the asymptotic behaviour, as $t \to \infty$, of large (resp. old) families, i.e. families with most frequent (resp. oldest) types, are developed in [CL12b].

In this article, we get explicit formulae for the expected site frequency spectrum $(\xi_k)_{1\leq k\leq n-1}$ of the sample under $\mathbb{P}_n^t$, $\mathbb{P}_n^{(\infty)}$, $\mathbb{P}_n^{(0)}$ and $\mathbb{P}_n^{(1)}$. According to Section 1.1, mutations are assumed to occur at constant rate θ on the lineages. Two different methods are used to obtain the expectation of the ξ_k. On the one hand, the similarity of the model with [Lam08] allows us to make use of a proof method developed in this article. Indeed, according to the results of Section 2, the framework used in [Lam08] covers our setting in the case of an infinite time of origin. On the other hand, for each k, $\mathbb{E}(\xi_k)$ can be expressed as a linear combination of the expectations of branching times [Wak09]. Although the first method could be used to prove all the results of this section, the second one provides very short proofs in the cases of an infinite time of origin and of a uniform prior. First under $\mathbb{P}_n^{(\infty)}$, the absence of a first moment for the time to the most recent common ancestor yields immediately the following result (Proposition 3.2).

Proposition. *For any $k \in \{1, \dots, n-1\}$, $\mathbb{E}_n^{(\infty)}(\xi_k)$ is infinite.*

Second, using the fact that the expected divergence times, under the Kingman coalescent model, and under the (suitably rescaled) critical birth-death process with uniform prior on its time of origin, are equal [Ger08], we deduce that the expected site frequency spectrum under $\mathbb{P}_n^{(0)}$ is that of a sample of the Kingman coalescent [Wak09, (4.20)] (Proposition 3.4).

Proposition. *For any $k \in \{1, \dots, n-1\}$, $\mathbb{E}_n^{(0)}(\xi_k) = n\theta/kp$.*

Finally, the formulas obtained in the remaining two cases are the following (Propositions 3.1 and 3.5).

Proposition. *For any $k \in \{1, \dots, n-1\}$, $t \in \mathbb{R}_+^*$, defining $\tau := pt$, we have*

$$\mathbb{E}_n^t(\xi_k) = \frac{\theta}{p}\left\{\frac{n-3k-1}{k} + \frac{(n-k-1)(k+1)}{k\tau}\right.$$
$$\left. + \frac{(1+\tau)^{k-1}}{\tau^{k+1}}\Big[2\tau^2 - (n-2k-1)2\tau - (n-k-1)(k+1)\Big]\left[\ln(1+\tau) - \sum_{i=1}^{k-1}\frac{1}{i}\left(\frac{\tau}{1+\tau}\right)^i\right]\right\}.$$

Proposition. *For any $k \in \{1, \dots, n-3\}$,*

$$\mathbb{E}_n^{(1)}(\xi_k) = \frac{\theta}{p}\,\frac{n(n-1)}{(n-k)(n-k-2)}\left[\frac{n+k-2}{k} - \frac{2(n-1)}{n-k-1}(\mathcal{H}_{n-1} - \mathcal{H}_k)\right],$$

where for any $k \in \mathbb{N}$, $\mathcal{H}_k = \sum_{j=1}^k j^{-1}$.

1.3.3 Convergence of genealogies for large sample sizes

We investigate in Section 4 the asymptotic behaviour of the coalescent point process π_n, as $n \to \infty$. We take inspiration from asymptotic results presented in [Pop04] and [AP05]. First, L. Popovic obtains in [Pop04] the convergence of the (suitably rescaled) coalescent point process of a critical birth-death process conditioned on its population size at time t towards a certain Poisson point measure on $(0,1) \times (0,t)$. Using this result, she then obtains with D. Aldous in

[AP05] a similar convergence for the model with uniform prior on the time of origin. Here we extend this to the cases of an infinite time of origin, and of a random time of origin with prior g_i, $i \in \mathbb{N}$.

Obtaining such asymptotic results requires to let the sampling parameter p depend on n in such a way that $p = n/\alpha$, with $\alpha > 0$. It ensures indeed that the expected number of sampled individuals is of the order of the sample size n. We then obtain the following convergences (Theorem 4.1).

Theorem. *Denote by π^t the Poisson point measure with intensity $\alpha \mathrm{dl}\, x^{-2}\mathrm{d}x\mathbb{1}_{(l,x)\in(0,1)\times(0,\alpha t)}$.*

a) *Under $\mathbb{P}_n^{(\infty)}$, the coalescent point process π_n converges in law, as $n \to \infty$, towards the Poisson point measure π with intensity measure $\alpha \mathrm{dl}\, x^{-2}\mathrm{d}x$ on $(0,1) \times \mathbb{R}_+^*$.*

b) *For any $i \in \mathbb{Z}_+$, under $\mathbb{P}_n^{(i)}$, the joint law of the time of origin, along with π_n, converges as $n \to \infty$ towards a pair $(T_{\mathrm{or}}^{(i)}, \pi^{(i)})$, such that $T_{\mathrm{or}}^{(i)}$ follows an inverse-gamma distribution with parameters $(i+1, \alpha)$, and conditional on $T_{\mathrm{or}}^{(i)} = t$, $\pi^{(i)}$ is distributed as π^t.*

The last result we obtain describes the links between the different random measures obtained in the limit. Let us order the atoms of our point processes w.r.t. their second coordinate. We prove that the random variable $T_{\mathrm{or}}^{(i)}$ is distributed as the $(i+1)$-th largest atom of the Poisson point process π, and we then deduce the following theorem (Theorem 4.4).

Theorem. *The point measure $\pi^{(i)}$ is distributed as the point process obtained from π by removing its $i+1$ largest atoms.*

In other words, genealogies with different priors can all be embedded in the same realization of the point measure π.

2 A universal distribution for the genealogy of a sample

Let us consider the model defined in Section 1.2 and specify some notation. Recall that the rescaled tree $\mathbb{T}$ is a critical birth-death tree with parameter N originating at time t, and that each extant individual in $\mathbb{T}$ is independently sampled with probability p/N.

We denote by $\mathcal{N}$ the number of extant individuals at present time in $\mathbb{T}$, and we label these individuals from 1 to $\mathcal{N}$, using the order defined in Section 0.1.4. In order to formalize the sampling process, we introduce a sequence $(I_j)_{j\geq 1}$ of random variables, such that $(I_1, I_2 - I_1, I_3 - I_2, ...)$ forms a sequence of i.i.d. geometric random variables with success probability p/N. Then for any j such that $I_j \leq \mathcal{N}$, I_j is the label of the j-th sampled individual in the extant population at present time (in the previously defined order). The conditioning on the sample size to be equal to n means thus conditioning on $\{I_n \leq \mathcal{N} < I_{n+1}\}$.

Let us now explain the link between the genealogy of the total extant population and the genealogy of the sample. Denote by $(H_i)_{1\leq i\leq \mathcal{N}-1}$ the sequence of node depths of the coalescent point process of the total extant population, i.e. for any $1 \leq i \leq \mathcal{N}-1$, H_i is the divergence time between individual i and individual $i+1$ in the rescaled tree $\mathbb{T}$. We know from [Lam10, Th.5.4] (or see Theorem 0.10) that for any $1 \leq i < j \leq \mathcal{N}$, the divergence time between individual i and

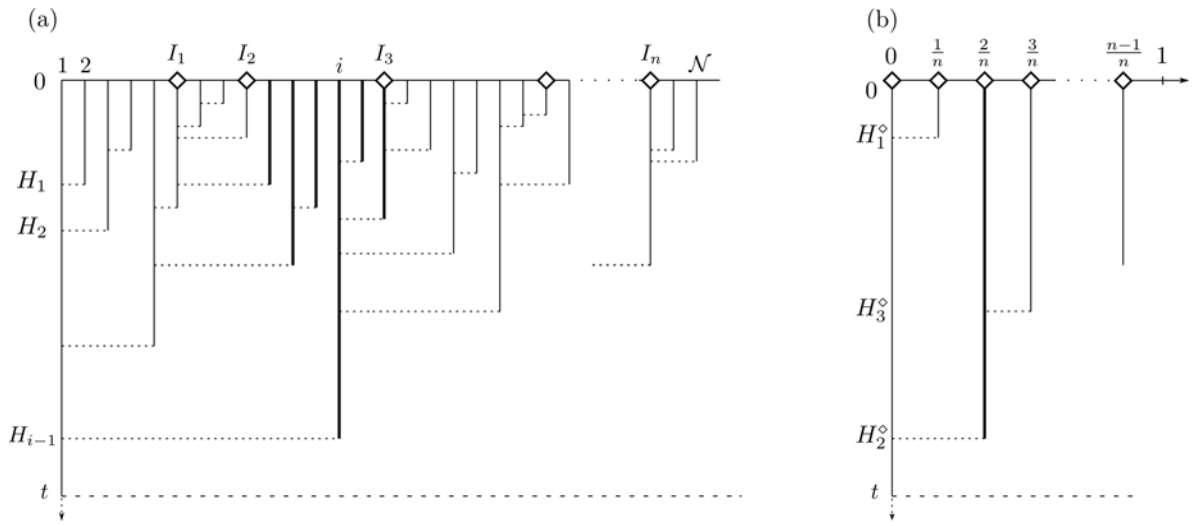

FIGURE 3 – (a)The coalescent point process at present time of a (rescaled) population originating at time t with n sampled individuals (symbolized by $\diamond$). The $\mathcal{N}$ vertical branches represent the sequence $(H_i)_{1\leq i\leq \mathcal{N}}$.
(b) The coalescent point process π_n of the sample represented in figure (a). The equality $H_2^\diamond = \max\{H_{I_2+1}, \ldots, H_{I_3}\}$ is illustrated by bold lines.

j is given by the maximum of the node depths $\{H_{i+1}, \ldots, H_j\}$. As a consequence, the divergence time $H_i^\diamond$ between individual I_i and individual I_{i+1} in $\mathbb{T}$, $1 \leq i \leq n-1$, is given by

$$H_i^\diamond = \max\{H_{I_i+1}, \ldots, H_{I_{i+1}}\}.$$

Finally we recall the definition of the point measure π_n :

$$\pi_n = \sum_{k=1}^{n-1} \delta_{(\frac{k}{n}, H_k^\diamond)}.$$

In the sequel we equally call « coalescent point process » of the sample, the measure π_n and the sequence $(H_i^\diamond)_{1\leq i\leq n-1}$. See Figure 3 for a graphical representation of the objects defined above.

The aim of this section is to characterize the law of the genealogy of the sample, under different assumptions on the time of origin. Section 2.1 establishes the distribution of π_n in the case of a fixed (possibly infinite) time of origin. In Section 2.2, we randomize the time of origin by giving it a prior distribution of the form $x \mapsto x^{-i}$, $i \in \mathbb{Z}_+$.

2.1 Fixed time of origin

We denote by $\mathbb{P}^t$ the law of the rescaled tree $\mathbb{T}$ originating at time t, and we recall that $\mathbb{P}_n^t$ denotes the law of $\mathbb{T}$ originating at time t and conditioned on having n sampled individuals at present time, i.e on $\{I_n \leq \mathcal{N} < I_{n+1}\}$. The following theorem specifies the law of the sample genealogy under $\mathbb{P}_n^t$.

Theorem 2.1. *Under $\mathbb{P}_n^t$, the coalescent point process $(H_i^\diamond)_{1\leq i\leq n-1}$ is a sequence of i.i.d. random variables with probability density function*

$$x \mapsto \frac{p}{(1+px)^2} \frac{1+pt}{pt} \mathbb{1}_{(0,t)}(x).$$

Remark 2.2. *According to [Pop04, Lem.3], the rescaled coalescent point process of the n sampled individuals is thus distributed as the coalescent point process of the population at time t of a critical branching process with rate p, conditioned on having n extant individuals at time t – or equivalently, as the coalescent point process of the population at time pt of a critical branching process with rate 1, conditioned on having n extant individuals at time pt, and then rescaled by a factor $1/p$.*

Remark 2.3. *It is interesting to note that the independence w.r.t. N of the law of π_n under $\mathbb{P}_n^t$ implies that the parameter N has only a scaling effect on the law of the genealogy. On the contrary, the parameters p and t both affect the branch lengths ratios, through the conditioning on the population size at a fixed time.*

We extend the theorem to the limiting case $t \to \infty$: recall that $\mathbb{P}_n^{(\infty)}(\mathbb{T} \in \cdot\,) = \lim_{t\to\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)$. We have the following statement.

Proposition 2.4. *Under $\mathbb{P}_n^{(\infty)}$, $(H_i^\diamond)_{1\le i\le n-1}$ is a sequence of i.i.d. random variables with density function $x \mapsto \frac{p}{(1+px)^2}\mathbb{1}_{\mathbb{R}_+}(x)$.*

Recall that for any $1 \le k \le n-1$, $T_{n,k}$ is defined as the k-th order statistic of the sequence $(H_i^\diamond)_{1\le i\le n-1}$. In particular, $T_{n,1}$ is the time to the most recent common ancestor of the sample. The following proposition provides the m-th moment of $T_{n,k}$ under $\mathbb{P}_n^{(\infty)}$.

Proposition 2.5. *For any $1 \le k \le n-1$ and $m \ge 1$, the m-th moment of $T_{n,k}$ under $\mathbb{P}_n^{(\infty)}$ is finite iff $m \le k-1$. Specifically, for $m \le k-1$,*

$$\mathbb{E}_n^{(\infty)}((T_{n,k})^m) = \frac{\binom{n-k+m-1}{m}}{p^m\binom{k-1}{m}}.$$

In particular, the time to the most recent common ancestor has infinite expectation under $\mathbb{P}_n^{(\infty)}$.

Proof of Proposition 2.5 :
Using the definition of $T_{n,k}$ as the k-th order statistic of the i.i.d. random variables $(H_i^\diamond)_{1\le i\le n-1}$ with density function $x \mapsto \frac{p}{(1+px)^2}\mathbb{1}_{\mathbb{R}_+}(x)$, along with [DN70, 2.1.6], we get that the density function of $T_{n,k}$ under $\mathbb{P}_n^{(\infty)}$ is $s \mapsto p(n-k)\binom{n-1}{n-k}\frac{(ps)^{n-k-1}}{(1+ps)^n}\mathbb{1}_{s\ge 0}$. Then

$$\mathbb{E}_n^{(\infty)}((T_{n,k})^m) = p^{-m}(n-k)\binom{n-1}{n-k}\int_0^\infty \frac{s^{n+m-k-1}}{(1+s)^n}\mathrm{d}s.$$

We conclude using Proposition A.2 in the Appendix. □

Proof of Theorem 2.1 :
For any $(t_1,\dots,t_{n-1}) \in (\mathbb{R}_+)^{n-1}$, write

$$\begin{aligned}&\mathbb{P}^t(H_1^\diamond \le t_1,\ \dots\ , H_{n-1}^\diamond \le t_{n-1},\ I_n \le \mathcal{N} \le I_{n+1} \,|\, \mathcal{N} \ge 1)\\ &= \sum_{k_0,\dots,k_n\ge 1} \mathbb{P}^t(H_1^\diamond \le t_1,\ \dots\ , H_{n-1}^\diamond \le t_{n-1},\ I_n \le \mathcal{N} \le I_{n+1},\ I_1 = k_0, \dots, I_n = k_0+\dots+k_n \,|\, \mathcal{N} \ge 1).\end{aligned}$$

Now recall from [Lam10, Th.5.4] that conditional on $\mathcal{N} \ge 1$, the sequence $(H_i)_{1\le i\le \mathcal{N}-1}$ is distributed as a sequence of i.i.d. random variables satisfying $\mathbb{P}^t(H_i \le u) = \frac{Nu}{1+Nu}$, stopped

at the first one exceeding t. Remembering that $H_i^\diamond = \max\{H_{I_i+1}, \ldots, H_{I_{i+1}}\}$, and from the definition of the sequence $(I_i)_{i\geq 1}$,

$$\begin{aligned}
&\mathbb{P}^t(H_1^\diamond \leq t_1,\ \ldots\ , H_{n-1}^\diamond \leq t_{n-1},\ I_n \leq \mathcal{N} \leq I_{n+1} \,|\, \mathcal{N} \geq 1) \\
&= \sum_{k_0,\ldots,k_n \geq 1} \Bigg(\prod_{i=0}^{n} \frac{p}{N} \Big(1 - \frac{p}{N}\Big)^{k_i - 1} \\
&\qquad\qquad \times \mathbb{P}^t\Big(\max_{1\leq i<l_0} H_i \leq t,\ \max_{l_0 \leq i < l_1} H_i \leq t_1, \ldots,\ \max_{l_{n-2}\leq i<l_{n-1}} H_i \leq t_{n-1},\ \max_{l_{n-1}\leq i<l_n} H_i > t\Big) \Bigg) \\
&= \sum_{k_0,\ldots,k_n \geq 1} \left[\prod_{i=0}^{n} \frac{p}{N} \Big(1 - \frac{p}{N}\Big)^{k_i - 1} \right] \left(\frac{Nt}{1+Nt} \right)^{k_0 - 1} \left[1 - \left(\frac{Nt}{1+Nt} \right)^{k_n} \right] \prod_{i=1}^{n-1} \left(\frac{N(t_i \wedge t)}{1 + N(t_i \wedge t)} \right)^{k_i},
\end{aligned}$$

where for any $0 \leq i \leq n$, $l_i := k_0 + \ldots + k_i$.

Now $\forall u \in \mathbb{R}_+$,

$$\sum_{k\geq 1} \frac{p}{N} \Big(1 - \frac{p}{N}\Big)^{k-1} \left(\frac{Nu}{1+Nu} \right)^k = \frac{pu}{1+pu},$$

and

$$\sum_{k_n \geq 1} \frac{p}{N} \Big(1 - \frac{p}{N}\Big)^{k_n - 1} \left(1 - \left(\frac{Nu}{1+Nu} \right)^{k_n} \right) = \frac{1}{1+pu}.$$

Thus we have

$$\begin{aligned}
&\mathbb{P}^t(H_1^\diamond \leq t_1,\ \ldots\ , H_{n-1}^\diamond \leq t_{n-1},\, I_n \leq \mathcal{N} \leq I_{n+1} \,|\, \mathcal{N} \geq 1) \\
&\qquad\qquad = \frac{1+Nt}{Nt} \, \frac{1}{1+pt} \, \frac{pt}{1+pt} \prod_{i=1}^{n-1} \frac{p(t_i \wedge t)}{1 + p(t_i \wedge t)}. \quad (1)
\end{aligned}$$

Finally, by taking $t_i = t$ for all $1 \leq i \leq n-1$ in (1), we get

$$\mathbb{P}^t(I_n \leq \mathcal{N} \leq I_{n+1} \,|\, \mathcal{N} \geq 1) = \frac{1+Nt}{Nt} \frac{1}{1+pt} \left(\frac{pt}{1+pt} \right)^n. \quad (2)$$

As a consequence, we have for any $(t_1, \ldots, t_{n-1}) \in (\mathbb{R}_+)^{n-1}$,

$$\mathbb{P}_n^t(H_1^\diamond \leq t_1,\ \ldots\ , H_{n-1}^\diamond \leq t_{n-1}) = \left(\frac{1+pt}{pt} \right)^{n-1} \prod_{i=1}^{n-1} \frac{p(t_i \wedge t)}{1 + p(t_i \wedge t)},$$

which leads to the announced result. □

2.2 Random time of origin

We now want to randomize the time of origin. To this aim, we give a (potentially improper) prior distribution to the time of origin in the model defined above. We investigate here priors with density function $g_i : u \mapsto u^{-i} \mathbb{1}_{\mathbb{R}_+^*}(u)$, $i \in \mathbb{Z}_+$. The case $i = 0$ (resp. $i = 1$) is usually referred to as uniform (resp. log-uniform) prior on $(0, \infty)$.

For any $0 \leq i < n$, recall that $\mathbb{P}_n^{(i)}$ denotes the law of the rescaled tree $\mathbb{T}$, with prior g_i on its time of origin, and conditioned on having n sampled individuals at present time :

$$\mathbb{P}_n^{(i)}(\mathbb{T} \in \cdot\,) = \frac{\int_0^{+\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)\mathbb{P}^t(I_n \leq \mathcal{N} \leq I_{n+1}) g_i(t)\,\mathrm{d}t}{\int_0^{+\infty} \mathbb{P}^t(I_n \leq \mathcal{N} \leq I_{n+1}) g_i(t)\,\mathrm{d}t}.$$

Note that we would have obtained the same distribution $\mathbb{P}_n^{(i)}$ if we had randomized the time of origin before having rescaled time in the process.

Proposition 2.6. *For any $0 \leq i < n$, the law of $\mathbb{T}$ under $\mathbb{P}_n^{(i)}$ is given by*

$$\mathbb{P}_n^{(i)}(\mathbb{T} \in \cdot\,) = \int_0^{+\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)\, h_n^{(i)}(t)\,\mathrm{d}t,$$

where

$$h_n^{(i)} : t \mapsto pn\binom{n-1}{i} \frac{(pt)^{n-i-1}}{(1+pt)^{n+1}} \mathbb{1}_{\mathbb{R}_+}(t),$$

i.e., the time of origin of $\mathbb{T}$ under $\mathbb{P}_n^{(i)}$ is a random variable T_{or} with posterior distribution characterized by its probability density function $h_n^{(i)}$.

Proof of Proposition 2.6 :
From (2) and from $\mathbb{P}^t(\mathcal{N} \geq 1) = (1+Nt)^{-1}$, we know that for all $t > 0$, $\mathbb{P}^t(I_n \leq \mathcal{N} \leq I_{n+1}) = \frac{1}{Nt}\frac{(pt)^n}{(1+pt)^{n+1}}$. Thus,

$$\int_0^{+\infty} \mathbb{P}^t(I_n \leq \mathcal{N} \leq I_{n+1}) g_i(t)\,\mathrm{d}t = \frac{p^i}{N}\int_0^{+\infty} \frac{(pt)^{n-i-1}}{(1+pt)^{n+1}}\, p\,\mathrm{d}t = \frac{p^i}{N}\frac{1}{(i+1)\binom{n}{i+1}} = \frac{p^i}{nN}\binom{n-1}{i}^{-1},$$

using Proposition A.2 in the Appendix. Finally by definition of $\mathbb{P}_n^{(i)}$,

$$\begin{aligned}\mathbb{P}_n^{(i)}(\mathbb{T} \in \cdot\,) &= \frac{Nn}{p^i}\binom{n-1}{i}\int_0^{+\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)\ \frac{p}{N}\frac{(pt)^{n-1}}{(1+pt)^{n+1}}\ \frac{\mathrm{d}t}{t^i} \\ &= \int_0^{+\infty} \mathbb{P}_n^t(\mathbb{T} \in \cdot\,)\ pn\binom{n-1}{i}\frac{(pt)^{n-i-1}}{(1+pt)^{n+1}}\,\mathrm{d}t,\end{aligned}$$

which gives the expected result. □

As a corollary, we have that the genealogy of the sample has the law of the genealogy of a birth-death process with fixed size :

Corollary 2.7. *For any $i \in \mathbb{Z}_+$, the rescaled coalescent point process π_n is distributed under $\mathbb{P}_n^{(i)}$ as the coalescent point process of a critical birth-death process with parameter p, with prior g_i on its time of origin, and conditioned on having n extant individuals at present time.*

Remark 2.8. *From the corollary it is easy to see that the sampling parameter p only has a scaling effect on time regarding the distribution of π_n under $\mathbb{P}_n^{(i)}$. This remains true under $\mathbb{P}_n^{(\infty)}$, but not under $\mathbb{P}_n^t$ because of the conditioning on the population size at time t (see Remark 2.3).*

Proof of Corollary 2.7 :
The probability for a critical birth-death process with parameter p of having n extant individuals

at time t is $\frac{(pt)^{n-1}}{(1+pt)^{n+1}}$ (see [AP05, (1)]), hence it differs from $\mathbb{P}^t(I_n \leq \mathcal{N} \leq I_{n+1})$ only by a factor p/N, and an easy adaptation of the calculations in the proof of Proposition 2.6 gives the expected result. □

Finally we study the moments of the divergence times $(T_{n,k})_{1\leq k\leq n-1}$. The following proposition states a necessary and sufficient condition for the existence of the m-th moment of $T_{n,k}$ under $\mathbb{P}_n^{(i)}$. In the case of a uniform prior ($i = 0$), we also recall the explicit formula established in [Ger08, Cor.2.2].

Proposition 2.9. *For any $0 \leq i < n$, $1 \leq k \leq n-1$ and $m \geq 1$, the m-th moment of $T_{n,k}$ under $\mathbb{P}_n^{(i)}$ is finite iff $m \leq k + i$.*
Besides, for any $1 \leq k \leq n-1$ and $m \leq k$,

$$\mathbb{E}_n^{(0)}((T_{n,k})^m) = \frac{\binom{n-k+m-1}{m}}{p^m\binom{k}{m}}.$$

Proof :
From Theorem 2.1, we know that under $\mathbb{P}_n^t$, the random variables $(H_i^\diamond)_{1\leq i\leq n-1}$ are i.i.d. Hence we obtain from [DN70, 2.1.6] that the random variable $T_{n,k}$, defined as the k-th order statistic of the sequence $(H_i^\diamond)_{1\leq i\leq n-1}$, has density function

$$f_{n,k}^t : s \mapsto p(n-k)\binom{n-1}{n-k}\frac{(ps)^{n-k-1}}{(1+ps)^n}\frac{(1+pt)^{n-k}}{(pt)^{n-1}}(pt-ps)^{k-1}\mathbb{1}_{s\leq t}$$

under $\mathbb{P}_n^t$. As a consequence, we have

$$\mathbb{E}_n^{(i)}((T_{n,k})^m) = \int_0^\infty s^m \left(\int_0^\infty f_{n,k}^t(s)h_n^{(i)}(t)\mathrm{d}t\right)\mathrm{d}s,$$

and then

$$\begin{aligned}\mathbb{E}_n^{(i)}((T_{n,k})^m) < \infty \;&\Leftrightarrow\; \int_0^\infty \frac{(ps)^{n-k-1+m}}{(1+ps)^n}\left(\int_{ps}^\infty \frac{(pt-ps)^{k-1}}{(pt)^i\,(1+pt)^{k+1}}\mathrm{d}t\right)\mathrm{d}s < \infty \\ &\Leftrightarrow\; \int_0^\infty \frac{s^{n-k-1+m}}{(1+s)^n}\left(\int_s^\infty \frac{(t-s)^{k-1}}{t^i\,(1+t)^{k+1}}\mathrm{d}t\right)\mathrm{d}s < \infty.\end{aligned}$$

Let us first characterize the integrability of the function $F : s \mapsto \frac{s^{n-k-1+m}}{(1+s)^n}\left(\int_s^\infty \frac{(t-s)^{k-1}}{t^i\,(1+t)^{k+1}}\mathrm{d}t\right)$ in the neighbourhood of $+\infty$. We prove here that $\int_s^\infty \frac{(t-s)^{k-1}}{t^i\,(1+t)^{k+1}}\mathrm{d}t \underset{s\to+\infty}{\sim} cs^{-i-1}$, where c is a (positive) constant w.r.t. s. Expanding $(t-s)^{k-1}$, we have

$$\int_s^\infty \frac{(t-s)^{k-1}}{t^i\,(1+t)^{k+1}}\mathrm{d}t = \sum_{j=0}^{k-1}(-s)^{k-1-j}\int_s^\infty \frac{\mathrm{d}t}{t^{i-j}(1+t)^{k+1}}.$$

Noting that for any $0 \leq j \leq k-1$,

$$\frac{1}{(k+i-j)(1+s)^{k+i-j}} \leq \int_s^\infty \frac{\mathrm{d}t}{t^{i-j}(1+t)^{k+1}} \leq \frac{1}{(k+i-j)s^{k+i-j}},$$

we obtain

$$\sum_{j=0}^{k-1}\frac{(-1)^{k-1-j}}{k+i-j}\binom{k-1}{j}\frac{s^{k+i-j}}{(1+s)^{k+i-j}} \leq s^{i+1}\int_s^\infty \frac{(t-s)^{k-1}}{t^i\,(1+t)^{k+1}}\mathrm{d}t \leq \sum_{j=0}^{k-1}\frac{(-1)^{k-1-j}}{k+i-j}\binom{k-1}{j},$$

Letting $s \to \infty$ leads to the announced equivalent. As a consequence, $F(s) \underset{+\infty}{\sim} cs^{m-k-i-2}$, and F is integrable in the neighbourhood of $+\infty$ iff $m-k-i \leq 0$.

On the other hand, in the case $m-k-i \leq 0$, the integrability of F on any compact set of $\mathbb{R}_+$ is clear. Thus $\mathbb{E}_n^{(i)}((T_{n,k})^m)$ is finite iff $m-k-i \leq 0$. □

3 Expected frequency spectrum

3.1 Mutation setting

Recall from Section 1 that we assume Poissonian mutations at rate $\theta \in \mathbb{R}_+$ on the lineages. We adopt the notation introduced in [Lam08], whose framework is very close to ours. Let $(\mathcal{P}_j)_{j\in\{0,\dots,n-1\}}$ be independent Poisson measures on $\mathbb{R}_+^*$ with parameter θ. For each j we denote the atom locations of $\mathcal{P}_j$ by $\ell_{j1} < \ell_{j2} < \dots$. The branch lengths $(H_0^\diamond, H_1^\diamond, \dots, H_{n-1}^\diamond)$, where we set $H_0^\diamond := T_{\mathrm{or}}$, characterize the genealogy of the n individuals (labeled accordingly from 0 to $n-1$) jointly with the foundation time of the population. Then the times ℓ_{jl} satisfying $\ell_{jl} < H_j^\diamond$ are interpreted as mutation events, and for all $k \in \{0,\dots,n-j-1\}$, individual $j+k$ bears mutation ℓ_{jl} if

$$\max\{H_{j+1}^\diamond, \dots, H_{j+k}^\diamond\} < \ell_{jl} < H_j^\diamond,$$

where $\max \varnothing = 0$ (see Figure 4). The first inequality expresses the fact that a mutation on branch j in the coalescent point process is carried by individual $j+k$ if the time at which it appears is greater than the divergence time of individuals j and $j+k$ (recall that time is running backwards). The second inequality means that all the values ℓ_{jl} that are greater than the j-th node depth $H_j^\diamond$ are not taken into account.

For any $k \in \{1,\dots,n-1\}$, recall that we denote by $(\xi_k)_{1\leq k\leq n-1}$ the site frequency spectrum of the sample, i.e. ξ_k is the number of mutations carried by k individuals among the n sampled individuals. The sum $S = \sum_{k=1}^{n-1} \xi_k$ is the so-called number of *polymorphic sites*, also known as *single nucleotide polymorphisms* in population genomics.

3.2 Results

In this section we give explicit formulae for the expected site frequency spectrum in the case of a fixed time of origin and in the case of a uniform or log-uniform prior on the time of origin. The proofs are based on two different methods, depending on the assumption on T_{or}, and are expanded in the next section.

3.2.1 Fixed (finite) time of origin

The expected site frequency spectrum of the n sampled individuals under $\mathbb{P}_n^t$ is given by

Proposition 3.1. *For any $k \in \{1,\dots,n-1\}$, $t \in \mathbb{R}_+^*$, defining $\tau := pt$, we have*

$$\mathbb{E}_n^t(\xi_k) = \frac{\theta}{p}\left\{\frac{n-3k-1}{k} + \frac{(n-k-1)(k+1)}{k\tau}\right.$$
$$\left. + \frac{(1+\tau)^{k-1}}{\tau^{k+1}}\Big[2\tau^2 - (n-2k-1)2\tau - (n-k-1)(k+1)\Big]\left[\ln(1+\tau) - \sum_{i=1}^{k-1}\frac{1}{i}\left(\frac{\tau}{1+\tau}\right)^i\right]\right\}.$$

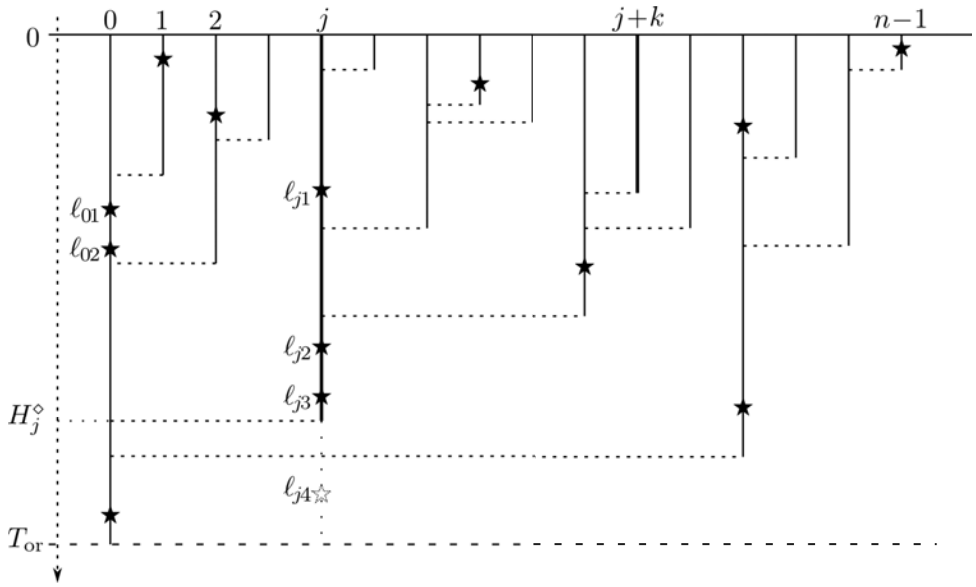

FIGURE 4 – The coalescent point process of a sample of size n, with mutations symbolized by stars. Mutations ℓ_{j2} and ℓ_{j3} are carried by individual $j+k$ while mutation ℓ_{j1} is not. Since $\ell_{j4} > H_j^\diamond$, it is not considered as a mutation event. Only mutations ℓ_{01}, ℓ_{02} and ℓ_{j1} are carried by two individuals, so that here $\xi_2 = 3$.

3.2.2 Infinite time of origin

The following two propositions are direct consequences of Proposition 3.1. However note that Proposition 3.2 can be proved independently from the formula provided by Proposition 3.1, as will be explained in Section 3.3.

Proposition 3.2. *For any $k \in \{1, \dots, n-1\}$, ξ_k has infinite expectation under $\mathbb{P}_n^{(\infty)}$.*

The infinite expectation of ξ_k under $\mathbb{P}_n^{(\infty)}$ leads to consider its renormalization by the expected number of polymorphic sites. The proposition below shows that letting the time of origin go to $+\infty$ flattens the renormalized expected frequency spectrum. A hint for this result is given in Section 3.3, while we prove it here by letting $t \to \infty$ in Proposition 3.1.

Proposition 3.3. *For any $k \in \{1, \dots, n-1\}$,*

$$\lim_{t\to\infty} \frac{\mathbb{E}_n^t(\xi_k)}{\mathbb{E}_n^t(S)} = \frac{1}{n-1}.$$

Proof :
Fix $k \in \{1, \dots, n-1\}$. One can easily see from Proposition 3.1 that as $t \to +\infty$,

$$\mathbb{E}_n^t(\xi_k) \sim 2\theta \ln(t) \quad \text{and} \quad \mathbb{E}_n^t(S) \sim 2\theta(n-1)\ln(t),$$

which leads to the result. □

3.2.3 Random time of origin

We provide explicit formulae for the expected frequency spectrum for two particular cases of priors : the uniform prior (case $i = 0$) and the log-uniform prior (case $i = 1$).

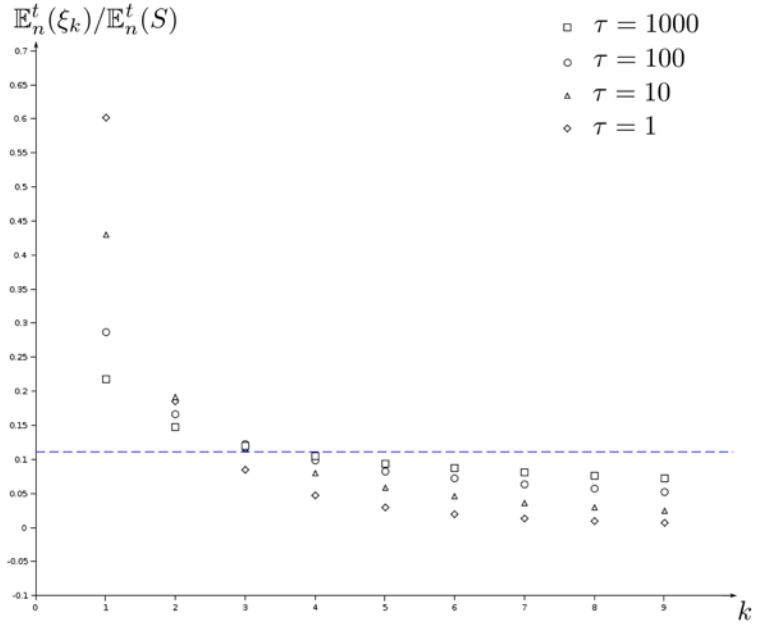

FIGURE 5 – The normalized expected site frequency spectrum of a sample of $n = 10$ individuals, under $\mathbb{P}_n^t$, for $\tau = pt \in \{1, 10, 100, 1000\}$. The horizontal dotted line has equation $y = 1/(n-1)$

Proposition 3.4. *For any* $k \in \{1, \dots, n-1\}$, $\mathbb{E}_n^{(0)}(\xi_k) = n\theta/pk$.

Proposition 3.5. *For any* $k \in \{1, \dots, n-3\}$,

$$\mathbb{E}_n^{(1)}(\xi_k) = \frac{\theta}{p} \frac{n(n-1)}{(n-k)(n-k-2)} \left[\frac{n+k-2}{k} - \frac{2(n-1)}{n-k-1}(\mathcal{H}_{n-1} - \mathcal{H}_k) \right],$$

where for any $k \in \mathbb{N}$, $\mathcal{H}_k = \sum_{j=1}^{k} j^{-1}$.

Remark 3.6. *The formulae obtained for* $\mathbb{E}_n^{(1)}(\xi_{n-2})$ *and* $\mathbb{E}_n^{(1)}(\xi_{n-3})$, *which we chose not to display here, involve non explicit integrals.*

Graphical representations of the expected frequency spectrum under $\mathbb{P}_n^t$, $\mathbb{P}_n^{(0)}$, $\mathbb{P}_n^{(1)}$ are provided in Figures 5 and 6.

3.3 Proofs

Depending on the assumption on T_{or}, two different methods can be used. The first one relies on an expression of the expected number of mutations carried by k individuals as a function of the expected coalescence times of the tree [Wak09, pp.103-105]. The second one decomposes its computation into the sum of the mutations present on lineage j, $1 \leq j \leq n-k$, carried by k individuals [Lam08]. Although the second one could be used to prove all the results of Section 3.2, the first one provides a very short proof in the cases of an infinite time of origin and of a uniform prior.

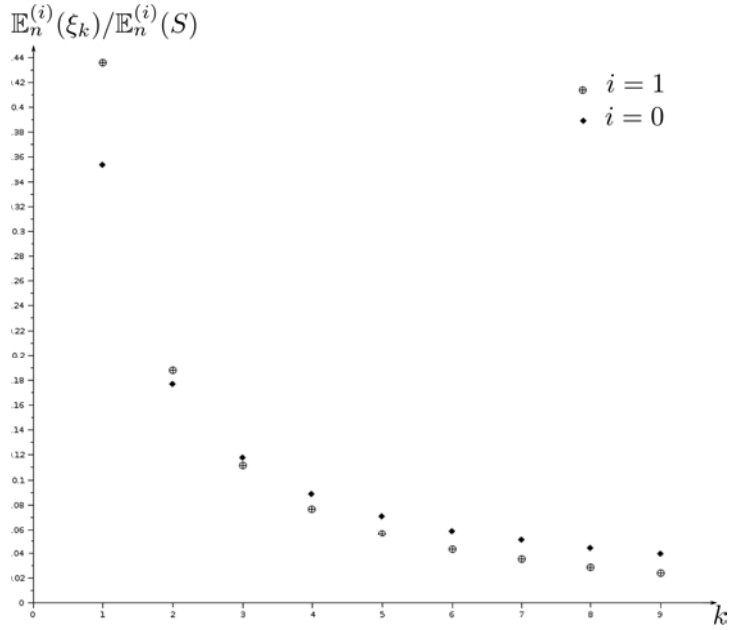

FIGURE 6 – The normalized expected site frequency spectrum of a sample of $n = 10$ individuals, under $\mathbb{P}_n^{(0)}$ and $\mathbb{P}_n^{(1)}$.

3.3.1 Infinite time of origin and uniform prior

We base our proof of Propositions 3.2 and 3.4 on Formula [Wak09, (4.22)], which gives for any $1 \leq k \leq n-1$ and any $i \in \mathbb{Z}_+ \cup \{\infty\}$

$$\mathbb{E}_n^{(i)}(\xi_k) = \theta \frac{2}{k} \binom{n-1}{k}^{-1} \sum_{j=2}^{n-k+1} \binom{j}{2} \binom{n-j}{k-1} \mathbb{E}_n^{(i)}(\hat{T}_{n,j}), \tag{3}$$

where $\hat{T}_{n,j} := T_{n,j} - T_{n,j-1}$ denotes the time elapsed between the $(j-1)$-th and the j-th coalescence.

When the time of origin is set to be infinite a.s., from Proposition 2.5 the expected time to the most recent common ancestor is infinite, which entails directly, along with (3), that $\mathbb{E}_n^{(\infty)}(\xi_k)$ is infinite for any $k \in \{1, \ldots, n-1\}$. From Equation (3) we can also give an intuitive explanation of the result of Proposition 3.3, which establishes that $\lim_{t\to\infty} \mathbb{E}_n^t(\xi_k)/\mathbb{E}_n^t(S) = 1/(n-1)$ for any $1 \leq k \leq n-1$. Indeed, using (3) to compute $\mathbb{E}_n^{(\infty)}(\xi_k)$, from Proposition 2.5 we know that $\mathbb{E}_n^{(\infty)}(\hat{T}_{n,2})$ is the only infinite contribution to $\mathbb{E}_n^{(\infty)}(\xi_k)$. This contribution is thus supported by the first order statistic $T_{n,1}$ of $(H_i^\diamond)_{1\leq i\leq n-1}$ (i.e. the largest divergence time in the coalescent point process). Conditional on $T_{n,1} = H_{i_0}^\diamond$, ξ_i is finite a.s. for any $i \neq n - i_0$. Now under $\mathbb{P}_n^{(\infty)}$, $(H_i^\diamond)_{1\leq i\leq n-1}$ is a sequence of i.i.d. random variables, so that the index i_0 is uniformly distributed in $\{1, \ldots, n-1\}$. This explains the independence of $\lim_{t\to\infty} \mathbb{E}_n^t(\xi_k)/\mathbb{E}_n^t(S)$ w.r.t. k.

In the case of a uniform prior on the time of origin, we use a comparison with the very documented Kingman coalescent model. Denote by $\mathbb{P}_{\mathrm{K}}$ the law of the genealogy of a sample of size n under the Kingman coalescent model with mutations at rate θ. First from [Ger08] we know that for any $j \in \{2, \ldots, n\}$, the inter-coalescence time $\hat{T}_{n,j}$ have proportional expectation under $\mathbb{P}_n^{(0)}$ and

under the Kingman coalescent model : $\mathbb{E}_n^{(0)}(\hat{T}_{n,j}) = \frac{n}{2p}\mathbb{E}_{\mathrm{K}}(\hat{T}_{n,j})$. Second, from [Wak09, (4.20)], for any $k \in \{1, \dots, n-1\}$, $\mathbb{E}_{\mathrm{K}}(\xi_k) = \frac{2\theta}{k}$. As a consequence, using (3) (which is also valid under $\mathbb{P}_{\mathrm{K}}$) we obtain for any $1 \leq k \leq n-1$, $\mathbb{E}_n^{(0)}(\xi_k) = \frac{n}{p}\frac{\theta}{k}$. This ends the proof of Proposition 3.4.

3.3.2 Fixed (finite) time of origin and log-uniform prior

When T_{or} is fixed (and finite), or in the case of a non uniform prior on T_{or} ($i \in \mathbb{N}$), the equality (3) does not lead to an explicit expression of the expected frequency spectrum. The formulae stated in Proposition 3.1 (case $T_{\mathrm{or}} = t \in \mathbb{R}_+^*$) and Proposition 3.5 (case of a log-uniform prior on T_{or}) are obtained using a method developed in [Lam08] (see proof of Theorem 2.3 for more details).

Proof of Proposition 3.1 :
Fix $t > 0$. Decomposing ξ_k into the sum of the number of mutations on the j-th branch carried by exactly k individuals, from [Lam08] (see proof of Theorem 2.3), we know that

$$\mathbb{E}_n^t(\xi_k) = \theta \sum_{j=0}^{n-k} \mathbb{E}_n^t\Big(\big(\min\{H_j^\diamond, H_{j+k}^\diamond\} - \max\{H_{j+1}^\diamond, \dots, H_{j+k-1}^\diamond\}\big)^+\Big), \tag{4}$$

where we have set $H_n^\diamond := +\infty$.

Two particular cases appear, namely $j = 0$, where $\min\{H_j^\diamond, H_{j+k}^\diamond\} = H_k^\diamond$ a.s., and $j + k = n$, where $\min\{H_j^\diamond, H_{j+k}^\diamond\} = H_j^\diamond$ a.s. Hence using the i.i.d. property of $(H_j^\diamond)_{1\leq j\leq n-1}$, it follows for any $1 \leq j \leq n-k-1$,

$$\begin{aligned} Q :&= \mathbb{E}_n^t\Big(\big(\min\{H_j^\diamond, H_{j+k}^\diamond\} - \max\{H_{j+1}^\diamond, \dots, H_{j+k-1}^\diamond\}\big)^+\Big) \\ &= \mathbb{E}_n^t \int_0^\infty \mathbb{1}_{\max\{H_{j+1}^\diamond,\dots,H_{j+k-1}^\diamond\} < x < \min\{H_j^\diamond, H_{j+k}^\diamond\}} \,\mathrm{d}x \\ &= \int_0^\infty \mathbb{P}_n^t(H_1^\diamond > x)^2 \, \mathbb{P}_n^t(H_1^\diamond < x)^{k-1} \,\mathrm{d}x, \end{aligned}$$

and similarly for $j \in \{0, n-k\}$,

$$R := \mathbb{E}_n^t\Big(\big(\min\{H_j^\diamond, H_{j+k}^\diamond\} - \max\{H_{j+1}^\diamond, \dots, H_{j+k-1}^\diamond\}\big)^+\Big) = \int_0^\infty \mathbb{P}_n^t(H_1^\diamond > x)\, \mathbb{P}_n^t(H_1^\diamond < x)^{k-1} \,\mathrm{d}x.$$

From Theorem 2.1, we know that $\mathbb{P}_n^t(H_1^\diamond < x) = \frac{p(x\wedge t)}{1+p(x\wedge t)}\frac{1+pt}{pt}$. This entails, after a change of variables, and recalling that we defined $\tau = pt$,

$$\begin{aligned} Q &= \frac{1}{p}\left(\frac{1+\tau}{\tau}\right)^{k-1} \int_0^\tau \left(\frac{x}{1+x}\right)^{k-1} \left(1 - \frac{x(1+\tau)}{\tau(1+x)}\right)^2 \mathrm{d}x \\ &= \frac{1}{p}\frac{(1+\tau)^{k-1}}{\tau^{k+1}}\big[\tau^2 I_{k+1,2}(\tau) - 2\tau I_{k+1,1}(\tau) + I_{k+1,0}(\tau)\big], \end{aligned}$$

and

$$\begin{aligned} R &= \frac{1}{p}\left(\frac{1+\tau}{\tau}\right)^{k-1} \int_0^\tau \left(\frac{x}{1+x}\right)^{k-1} \left(1 - \frac{x(1+\tau)}{\tau(1+x)}\right) \mathrm{d}x \\ &= \frac{1}{p}\frac{(1+\tau)^{k-1}}{\tau^{k+1}}\big[\tau^2 I_{k,1}(\tau) - \tau I_{k,0}(\tau)\big], \end{aligned}$$

where for any $u \in \mathbb{R}_+^*$, $k \in \mathbb{Z}_+$, $l \in \mathbb{Z}$, $I_{k,l}(u) := \int_0^u \frac{x^{k-l}}{(1+x)^k} \mathrm{d}x$. Using Equation (4), this leads to

$$\mathbb{E}_n^t(\xi_k) = \frac{\theta}{p} \frac{(1+\tau)^{k-1}}{\tau^{k+1}} \Big[(n-k-1)\big(\tau^2 I_{k+1,2}(\tau) - 2\tau I_{k+1,1}(\tau) + I_{k+1,0}(\tau)\big) + 2\big(\tau^2 I_{k,1}(\tau) - \tau I_{k,0}(\tau)\big)\Big]$$

Finally, using the formulae provided by Proposition A.1 for $I_{k,l}$, $l \in \{0,1,2\}$, we finally get after some rearrangements

$$\begin{aligned}\mathbb{E}_n^t(\xi_k) = \frac{\theta}{p} \frac{(1+\tau)^{k-1}}{\tau^{k+1}} \Bigg[& \ln(1+\tau)\Big(2\tau^2 - 2(n-2k-1)\tau - (k+1)(n-k-1))\Big) \\ & - 2\tau^2 + \tau(n-k-1) + \frac{n-k-1}{k} \frac{\tau^{k+2}}{(1+\tau)^k} + (-1)^{k-1} \frac{n-k-1}{k} \left(1 - \frac{1}{(1+\tau)^k}\right)(2\tau+1) \\ & + \sum_{j=1}^{k-1} \binom{k-1}{j} \frac{(-1)^j}{j} \left(1 - \frac{1}{(1+\tau)^j}\right) \left(2\tau^2 + \frac{2\tau k}{j+1} - (n-k-1)\left(2\tau + \frac{k+1}{j+1}\right)\frac{k}{k-j}\right) \Bigg]. \end{aligned} \tag{5}$$

To obtain the final form, we decompose the sum in the r.h.s. as follows :
First define, for $x \in \mathbb{R}$ and $k \in \mathbb{N}$

$$\begin{array}{ll} \phi_{1,k}(x) := \sum_{j=1}^k \binom{k}{j} \frac{x^j}{j}, & \phi_{2,k}(x) := \sum_{j=1}^k \binom{k}{j} \frac{x^{j+1}}{j(j+1)} \\ \psi_{1,k}(x) := \sum_{j=1}^k \binom{k}{j} \frac{x^j}{j(k-j)}, & \psi_{2,k}(x) := \sum_{j=1}^k \binom{k}{j} \frac{x^{j+1}}{j(j+1)(k-j)}. \end{array}$$

Then we have

$$\begin{aligned} S :=& \sum_{j=1}^{k-1} \binom{k-1}{j} \frac{(-1)^j}{j} \left(1 - \frac{1}{(1+\tau)^j}\right) \left(2\tau^2 + \frac{2\tau k}{j+1} - (n-k-1)\left(2\tau + \frac{k+1}{j+1}\right)\frac{k}{k-j}\right) \\ =& \ 2\tau^2\big(\phi_{1,k-1}(-1) - \phi_{1,k-1}(-(1+\tau)^{-1})\big) \\ & - 2\tau k\big(\phi_{2,k-1}(-1) - (1+\tau)\phi_{2,k-1}(-(1+\tau)^{-1})\big) \\ & - (n-k-1)2\tau k\big(\psi_{1,k-1}(-1) - \psi_{1,k-1}(-(1+\tau)^{-1})\big) \\ & + (n-k-1)(k+1)k\big(\psi_{2,k-1}(-1) - (1+\tau)\psi_{2,k-1}(-(1+\tau)^{-1})\big). \end{aligned}$$

Let us now reexpress the functions $\phi_{1,k}$, $\phi_{2,k}$, $\psi_{1,k}$ and $\psi_{2,k}$. Fix $x \in \mathbb{R}$ and $k \in \mathbb{N}$. The function $\phi_{1,k}$ is differentiable at x and we have $\phi'_{1,k}(x) = \sum_{j=1}^k \binom{k}{j} x^{j-1} = x^{-1}\big[(1+x)^k - 1\big]$. This leads by simple integration calculus to $\phi_{1,k}(x) = \sum_{j=1}^k \frac{(1+x)^j}{j} - \mathcal{H}_k$, where $\mathcal{H}_k = \sum_{j=1}^k j^{-1}$. Then noting that $\phi'_{2,k}(x) = \phi_{1,k}(x)$, we obtain $\phi_{2,k}(x) = \sum_{j=1}^k \frac{(1+x)^{j+1}}{j(j+1)} - x\mathcal{H}_k + \frac{1}{k+1} - 1$. Finally, it is easy to show that $\psi_{1,k}(x) = \frac{1}{k+1}\big(\phi_{1,k+1}(x) - \frac{x^{k+1}}{k+1}\big)$ and $\psi_{2,k}(x) = \frac{1}{k+1}\big(\phi_{2,k+1}(x) - \frac{x^{k+2}}{(k+1)(k+2)}\big)$. This

yields

$$
\begin{aligned}
S = &-2\tau^2 \sum_{i=1}^{k-1} \frac{1}{i} \left(\frac{\tau}{1+\tau}\right)^i \\
&+ 2\tau k \left[\frac{k-1}{k}\tau + \tau \sum_{i=1}^{k-1} \frac{1}{i(i+1)} \left(\frac{\tau}{1+\tau}\right)^i\right] \\
&+ 2(n-k-1)\tau \left[\sum_{i=1}^{k} \frac{1}{i} \left(\frac{\tau}{1+\tau}\right)^i + \frac{(-1)^k}{k}\left(1 - \frac{1}{(1+\tau)^k}\right)\right] \\
&+ (n-k-1)(k+1) \left[\frac{k}{k+1}\tau - \tau \sum_{i=1}^{k} \frac{1}{i(i+1)} \left(\frac{\tau}{1+\tau}\right)^i + \frac{(-1)^k}{k(k+1)}\left(1 - \frac{1}{(1+\tau)^k}\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= (n-k-1)k\tau - 2(k-1)\tau^2 + \frac{n-k-1}{k}\frac{\tau^{k+1}}{(1+\tau)^k} + (-1)^k \frac{n-k-1}{k}\left(1 - \frac{1}{(1+\tau)^k}\right)(1+2\tau) \\
&+ (2\tau(n-k-1) - 2\tau^2) \sum_{i=1}^{k-1} \frac{1}{i} \left(\frac{\tau}{1+\tau}\right)^i + (2\tau^2 k - (n-k-1)(k+1)\tau) \sum_{i=1}^{k-1} \frac{1}{i(i+1)} \left(\frac{\tau}{1+\tau}\right)^i \\
&= 2\tau^2 - \tau(n-k-1) + \frac{n-k-1}{k}\frac{\tau^{k+1}}{(1+\tau)^k} + (-1)^k \frac{n-k-1}{k}\left(1 - \frac{1}{(1+\tau)^k}\right)(2\tau+1) \\
&\qquad - \frac{1}{k}\left(\frac{\tau}{1+\tau}\right)^{k-1} \big(2k\tau^2 - (n-k-1)(k+1)\tau\big) \\
&\qquad + \big(2(n-2k-1)\tau - 2\tau^2 + (k+1)(n-k-1)\big) \sum_{i=1}^{k-1} \frac{1}{i} \left(\frac{\tau}{1+\tau}\right)^i,
\end{aligned}
$$

where the last equality was obtained by writing $\frac{1}{i(i+1)} = \frac{1}{i} - \frac{1}{i+1}$. It suffices now to reinject this formula into equation (5) to obtain the announced result. □

Proof of Proposition 3.5 :
Reasoning as in the proof of Proposition 3.1, we express $\mathbb{E}_n^{(1)}(\xi_k)$ as

$$
\mathbb{E}_n^{(1)}(\xi_k) = \theta \sum_{j=0}^{n-k} \mathbb{E}_n^{(1)}\Big(\big(\min\{H_j^\diamond, H_{j+k}^\diamond\} - \max\{H_{j+1}^\diamond, \dots, H_{j+k-1}^\diamond\}\big)^+\Big), \tag{6}
$$

with for any $1 \le j \le n-k-1$,

$$
\begin{aligned}
Q :&= \mathbb{E}_n^{(1)}\Big(\big(\min\{H_j^\diamond, H_{j+k}^\diamond\} - \max\{H_{j+1}^\diamond, \dots, H_{j+k-1}^\diamond\}\big)^+\Big) \\
&= \int_0^\infty \left(\int_0^\infty h_n^{(1)}(\tau)\, \mathbb{P}_n^t(H_1^\diamond > x)^2\, \mathbb{P}_n^t(H_1^\diamond < x)^{k-1}\, \mathrm{d}\tau\right) \mathrm{d}x,
\end{aligned}
$$

and for $j \in \{0, n-k\}$,

$$
\begin{aligned}
R :&= \mathbb{E}_n^{(1)}\Big(\big(\min\{H_j^\diamond, H_{j+k}^\diamond\} - \max\{H_{j+1}^\diamond, \dots, H_{j+k-1}^\diamond\}\big)^+\Big) \\
&= \int_0^\infty \left(\int_0^\infty h_n^{(1)}(\tau)\, \mathbb{P}_n^t(H_1^\diamond > x)\, \mathbb{P}_n^t(H_1^\diamond < x)^{k-1}\, \mathrm{d}\tau\right) \mathrm{d}x.
\end{aligned}
$$

From Theorem 2.1, we know that $\mathbb{P}_n^t(H_1^\diamond < x) = \frac{p(x\wedge t)}{1+p(x\wedge t)}\frac{1+pt}{pt}$, and from Proposition 2.6, for all $t \geq 0$, $h_n^{(1)}(t) = pn(n-1)\frac{(pt)^{n-2}}{(1+pt)^{n+1}}$. After a change of variables, this leads to

$$\begin{aligned}
Q &= \frac{1}{p}n(n-1)\int_0^\infty \left(\frac{x}{1+x}\right)^{k-1}\left(\int_x^\infty \frac{t^{n-k-1}}{(1+t)^{n-k+2}}\left(1-\frac{x(1+t)}{t(1+x)}\right)^2 \mathrm{d}t\right)\mathrm{d}x \\
&= \frac{1}{p}n(n-1)\int_0^\infty \frac{x^{k-1}}{(1+x)^{k+1}}\big[J_{n-k+2,3}(x) - 2x\, J_{n-k+2,4}(x) + x^2 J_{n-k+2,5}(x)\big]\mathrm{d}x, \\
R &= \frac{1}{p}n(n-1)\int_0^\infty \left(\frac{x}{1+x}\right)^{k-1}\left(\int_x^\infty \frac{t^{n-k-1}}{(1+t)^{n-k+2}}\left(1-\frac{x(1+t)}{t(1+x)}\right) \mathrm{d}t\right)\mathrm{d}x \\
&= \frac{1}{p}n(n-1)\int_0^\infty \frac{x^{k-1}}{(1+x)^{k}}\big[J_{n-k+2,3}(x) - x\, J_{n-k+2,4}(x)\big]\mathrm{d}x,
\end{aligned}$$

where for any integers $k \geq l \geq 2$ and for any positive real number x, $J_{k,l}(x) := \int_x^\infty \frac{u^{k-l}}{(1+u)^k}\mathrm{d}u$.

Now using (9) in Proposition A.2 to express the integrals $J_{k,l}$ in R and Q, and using again Proposition A.2 to calculate the remaining integrals, we obtain for any $k \geq n-3$,

$$\begin{aligned}
Q = \frac{1}{p}\frac{2n(n-1)}{(n-k)(n-k+1)}\Bigg[&\sum_{j=0}^{n-k-1}\frac{j+1}{(j+k)(j+k+1)(j+k+2)} \\
&- \frac{2}{(n-k-1)}\sum_{j=0}^{n-k-2}\frac{(j+1)(j+2)}{(j+k+1)(j+k+2)(j+k+3)} \\
&+ \frac{1}{(n-k-1)(n-k-2)}\sum_{j=0}^{n-k-3}\frac{(j+1)(j+2)(j+3)}{(j+k+2)(j+k+3)(j+k+4)}\Bigg],
\end{aligned}$$

$$R = \frac{1}{p}\frac{n(n-1)}{(n-k)(n-k+1)}\left[\sum_{j=0}^{n-k-1}\frac{j+1}{(j+k)(j+k+1)} + \frac{1}{n-k-1}\sum_{j=0}^{n-k-2}\frac{(j+1)(j+2)}{(j+k+1)(j+k+2)}\right].$$

Finally, using partial fraction decompositions to calculate the sums in the expressions of Q and R,

$$\begin{aligned}
Q &= \frac{1}{p}\frac{n(n-1)}{(n-k)(n-k+1)}\left[\frac{1}{k} + \frac{6}{n-k-2} - \frac{2(2n+k-1)}{(n-k-1)(n-k-2)}(\mathcal{H}_{n-1} - \mathcal{H}_k)\right], \\
R &= \frac{1}{p}\frac{n(n-1)}{(n-k-1)(n-k+1)}\left[\frac{n+k-1}{n-k}(\mathcal{H}_{n-1} - \mathcal{H}_{k-1}) - 2\right].
\end{aligned}$$

Reinjecting these expressions into equation (6) leads to

$$\mathbb{E}_n^{(1)}(\xi_k) = \frac{\theta}{p}[(n-k-1)Q + 2R] = \frac{\theta}{p}\frac{n(n-1)}{(n-k)(n-k-2)}\left[\frac{n+k-2}{k} - \frac{2(n-1)}{n-k-1}(\mathcal{H}_{n-1} - \mathcal{H}_k)\right],$$

for any $1 \leq k \leq n-3$, which ends the proof. □

4 Convergence of genealogies in the large sample asymptotic

In this section we provide convergence results for the distribution of the suitably rescaled genealogy of a sample of size n, as $n \to \infty$. Obtaining such asymptotic results requires an additional assumption on the sampling probability : we assume that the sampling parameter p depends on n in such a way that $p = n/\alpha$, where $\alpha \in \mathbb{R}_+^*$. This assumption arises naturally, since it ensures that the expected number of sampled individuals is of order n. Besides, according to Remark 2.8, note that the parameter α will only have a scaling effect on time.

In the sequel, the symbol $\stackrel{\mathcal{L}}{=}$ means an equality in law, and for any $n > i \geq 0$, $\mathcal{L}(\,\cdot\,,\ \mathbb{P}_n^{(i)})$ refers to the distribution of a random variable or a process under $\mathbb{P}_n^{(i)}$. Finally, $\Rightarrow$ denotes the convergence in distribution. Recall from Section 0.1.2 that, if (γ_n) is a sequence of random measures on $\mathbb{R}^d$ and γ a simple point process on $\mathbb{R}^d$, $\gamma_n \Rightarrow \gamma$ iff $\gamma_n(B) \Rightarrow \gamma(B)$ for any compact set B such that $\gamma(\partial B) = 0$, where ∂B denotes the boundary of B.

4.1 Results

Convergence of genealogies

First define, for any $t > 0$, π^t (resp π) as the Poisson point measure on $(0,1) \times (0, \alpha t)$ (resp. $(0,1) \times \mathbb{R}_+^*$) with intensity $\alpha \mathrm{d}l\, x^{-2} \mathrm{d}x \mathbb{1}_{(l,x) \in (0,1) \times (0,\alpha t)}$ (resp. $\alpha \mathrm{d}l\, x^{-2} \mathrm{d}x \mathbb{1}_{(l,x) \in (0,1) \times \mathbb{R}_+^*}$).

Let $(\rho_i)_{i \geq 0}$ be a sequence of i.i.d. exponential random variables with parameter $1/\alpha$, and define for all $i \geq 0$ the inverse-gamma random variable $\mathrm{e}_i := (\rho_0 + \ldots + \rho_i)^{-1}$. Then for $i \in \mathbb{Z}_+$, define the pair $(\pi^{(i)}, T_{\mathrm{or}}^{(i)})$, where $T_{\mathrm{or}}^{(i)}$ is a positive random variable, and $\pi^{(i)}$ is a Cox process $\pi^{(i)}$, as

$$\mathbb{P}(T_{\mathrm{or}}^{(i)} \in \mathrm{d}t,\ \pi^{(i)} \in \cdot) = \mathbb{P}(\mathrm{e}_i \in \mathrm{d}t) \mathbb{P}(\pi^t \in \cdot).$$

In particular, conditional on $T_{\mathrm{or}}^{(i)} = t$, $\pi^{(i)}$ has the law of the Poisson point measure π^t.

The first theorem states the convergence in distribution of the random measure π_n under $\mathbb{P}_n^{(i)}$, $i \in \mathbb{Z}_+ \cup \{\infty\}$. This result is a generalization of Corollary 2 in [AP05], which provides convergence in distribution of π_n under $\mathbb{P}_n^{(0)}$ towards $\pi^{(0)}$. The proof of this convergence, as well as the proof of the generalization we propose, mainly rely on the convergence of π_n under $\mathbb{P}_n^t$ towards π^t, which is established in Theorem 5 in [Pop04].

Theorem 4.1. *We have the following convergences in distribution as $n \to \infty$:*

a) $$\mathcal{L}(\pi_n,\ \mathbb{P}_n^{(\infty)}) \Rightarrow \pi,$$

b) *and for any $i \geq 0$,* $$\mathcal{L}((\pi_n, T_{\mathrm{or}}),\ \mathbb{P}_n^{(i)}) \Rightarrow (\pi^{(i)}, T_{\mathrm{or}}^{(i)}).$$

As a corollary of this theorem, we state the finite dimensional convergence of the divergence times of π_n under $\mathbb{P}_n^{(i)}$, $i \in \mathbb{Z}_+ \cup \{\infty\}$. We denote by $(T_k)_{k \geq 1}$ (resp. $(T_k^{(i)})_{k \geq 1}$) the decreasing reordering of the second coordinates of the atoms of π (resp. $\pi^{(i)}$).

Corollary 4.2. *Fix $k \in \mathbb{N}$. We have the following convergences in distribution as $n \to \infty$:*

a) $$\mathcal{L}((T_{n,1}, \ldots, T_{n,k}),\ \mathbb{P}_n^{(\infty)}) \Rightarrow (T_1, \ldots, T_k),$$

b) *and for any $i \geq 0$,* $$\mathcal{L}((T_{\mathrm{or}}, T_{n,1}, \ldots, T_{n,k}),\ \mathbb{P}_n^{(i)}) \Rightarrow (T_{\mathrm{or}}^{(i)}, T_1^{(i)}, \ldots, T_k^{(i)}).$$

Besides, the limiting distributions appearing in Corollary 4.2 are specified in the following proposition.

Proposition 4.3.

a) *For any $k \in \mathbb{N}$, the k-tuple $(T_1, \dots, T_k)$ is distributed as $(\mathrm{e}_0, \dots, \mathrm{e}_{k-1})$.*

b) *For any $i \in \mathbb{Z}_+$, $k \in \mathbb{N}$, the $k+1$-tuple $(T_{\mathrm{or}}^{(i)}, T_1^{(i)}, \dots, T_k^{(i)})$ is distributed as $(\mathrm{e}_i, \dots, \mathrm{e}_{i+k})$.*

The last theorem describes the links between the different random measures obtained in the limit, in Theorem 4.1. Before stating this result, let us clarify some definition. Consider μ any random measure among π, $\pi^{(i)}$ $(i \in \mathbb{Z}_+)$. Conditional on $\mu = \sum_{t \in A} \delta_{(t,y_t)}$, where $A \subset [0,1]$ is a countable set, denoting by (u, y_u) its largest atom, where we refer to the order w.r.t. the second coordinate, we define the random measure $\sum_{t \in A \setminus \{u\}} \delta_{(t,y_t)}$ as the random measure obtained from μ by removing its largest atom.

Proposition 4.3 establishes in particular that for any $i \in \mathbb{Z}_+$, the time of origin $T_{\mathrm{or}}^{(i)}$ is distributed as the $(i+1)$-th largest atom of the random measure π. The following statement is a direct consequence of this result.

Theorem 4.4. *For any $i \in \mathbb{Z}_+$, the measure $\pi^{(i)}$ has the distribution of the random measure obtained from π by removing its $i+1$ largest atoms. In particular, for any $i \in \mathbb{N}$, the measure $\pi^{(i)}$ has the distribution of the random measure obtained from $\pi^{(i-1)}$ by removing its largest atom.*

As a conclusion, in the limit $n \to \infty$, genealogies with different priors on the time of origin can all be embedded in the same realization of the measure π : a realization of the limiting coalescent point process with given prior can be obtained by removing from a realization of π a given number of its largest atoms.

Convergence of the expected site frequency spectrum

Recall that mutations are assumed to occur at rate θ on the lineages. We deduce the following proposition from the results of Section 3.2.

Proposition 4.5. *For any $t \in \mathbb{R}_+^*$ and any $i \in \{0,1\}$, for any $k \in \mathbb{N}$ we have*

$$\lim_{n\to\infty} \mathbb{E}_n^t(\xi_k) = \alpha\theta/k \quad \textit{and} \quad \lim_{n\to\infty} \mathbb{E}_n^{(i)}(\xi_k) = \alpha\theta/k.$$

In other words, under $\mathbb{P}_n^t$, $\mathbb{P}_n^{(0)}$ and $\mathbb{P}_n^{(1)}$, the expected site frequency spectrum of the sample converges, as the sample size gets large, towards the expected frequency spectrum of the Kingman coalescent [Wak09, (4.20)].

4.2 Proofs

To begin with, we state the convergence, as $n \to \infty$, of the posterior distribution of the time of origin T_{or} under $\mathbb{P}_n^{(i)}$. This result is essential to obtain other convergence results under $\mathbb{P}_n^{(i)}$, since the posterior density function $h_n^{(i)}$ of T_{or} is directly involved in the definition of the law $\mathbb{P}_n^{(i)}$.

Proposition 4.6. *For any $i \in \mathbb{Z}_+$, we have the following convergence in law*

$$\mathcal{L}(T_{\mathrm{or}},\ \mathbb{P}_n^{(i)}) \Rightarrow T_{\mathrm{or}}^{(i)}.$$

Lemma 4.7. *For any $i \in \mathbb{Z}_+$, the random variable e_i has density function $h^{(i)} : t \mapsto \frac{\alpha^{i+1} e^{-\alpha/t}}{i!\, t^{i+2}} \mathbb{1}_{t>0}$ (i.e. e_i follows an inverse-gamma distribution with parameters $(i+1, \alpha)$).*

Proof :
Fix $i \in \mathbb{Z}_+$. The random variable e_i is the inverse of the sum of $i+1$ independent exponential variables with parameter α^{-1}, i.e. the inverse of a Gamma variable with parameters $(i+1, \alpha^{-1})$. From the known density function $t \mapsto \frac{\alpha^{i+1} t^i e^{-\alpha t}}{i!} \mathbb{1}_{t>0}$ of a $\Gamma(i+1, \alpha^{-1})$-variable, we deduce that e_i has density function $t \mapsto \frac{\alpha^{i+1} e^{-\alpha/t}}{i!\, t^{i+2}} \mathbb{1}_{t>0}$. □

Proof of Proposition 4.6 :
Recall first that by definition, $T_{\mathrm{or}}^{(i)}$ is distributed as e_i, and as a consequence, has density function $h^{(i)}$. From Proposition 2.6, recalling that $p = n/\alpha$, the density function of T_{or} under $\mathbb{P}_n^{(i)}$ is given by : for all $t > 0$,

$$h_n^{(i)}(t) = \frac{n^2}{\alpha\, i!}(n-1)\dots(n-i)\left(\frac{nt/\alpha}{1+nt/\alpha}\right)^{n+1} \frac{1}{(nt/\alpha)^{i+2}} = \frac{(n-1)\dots(n-i)}{n^i} \frac{e^{-(n+1)\ln(1+\frac{\alpha}{nt})}}{i!\ \alpha\ (t/\alpha)^{i+2}}, \tag{7}$$

and hence for all $t > 0$

$$h_n^{(i)}(t) \underset{n\to\infty}{\to} \frac{\alpha^{i+1} e^{-\alpha/t}}{i!\, t^{i+2}} = h^{(i)}(t),$$

and the convergence of the density functions $(h_n^{(i)})$ towards $h^{(i)}$ ensures the convergence in law under $\mathbb{P}_n^{(i)}$ of T_{or} towards $T_{\mathrm{or}}^{(i)}$. □

To prove Theorem 4.1, we first recall Theorem 5 of [Pop04], which can be stated as follows.

Lemma 4.8. *For any $t > 0$, as $n \to \infty$, $\mathcal{L}(\pi_n,\ \mathbb{P}_n^t) \Rightarrow \pi^t$.*

Proof of Theorem 4.1 :
a) From Proposition 2.4, under $\mathbb{P}_n^{(\infty)}$ the random measure π_n is a simple point process with intensity $\sum_{i=1}^{n-1} \delta_{\{i/n\}}(\mathrm{d}l)\, \frac{n\mathrm{d}x}{\alpha(1+nx/\alpha)^2}$. As $n \to \infty$, this intensity measure converges weakly towards $\alpha \mathrm{d}l\, x^{-2}\mathrm{d}x \mathbb{1}_{(l,x)\in(0,1)\times(0,+\infty)}$, which is the intensity measure of the Poisson process π. From Corollary 0.4 (or [Kal02, Th.16.18]), this is sufficient to prove the convergence in distribution, under $\mathbb{P}_n^{(\infty)}$, of π_n towards π.

b) Fix $i \in \mathbb{Z}_+$. For any compact set A of $[0,1]\times\mathbb{R}_+$, B Borel set of $\mathbb{R}_+$ of zero Lebesgue measure boundary, and $k \in \mathbb{Z}_+$, we have

$$\mathbb{P}_n^{(i)}(\pi_n(A) = k,\ T_{\mathrm{or}} \in B) = \int_B \mathbb{P}_n^t(\pi_n(A) = k)\, h_n^{(i)}(t)\, \mathrm{d}t.$$

In order to apply the dominated convergence theorem, we first remark that for all $t > 0$, for any $n > i$,

$$h_n^{(i)}(t) \le f^{(i)}(t) := \frac{\alpha^{i+1}}{i!\, t^{i+2}}, \tag{8}$$

as can easily be seen from (7). Besides, studying the variations of $h_n^{(i)}$ yields in particular that $h_n^{(i)}$ is a nonnegative function that increases on $\left(0, \alpha\frac{n-i-1}{n(i+2)}\right)$. Now there exists $\beta > 0$ such that for n large enough, $\alpha\frac{n-i-1}{n(i+2)} \ge \beta$. Finally we have from (8) that for any n large enough and for all $t > 0$,

$$|h_n^{(i)}(t)| \le f^{(i)}(\beta)\mathbb{1}_{t\le\beta} + f^{(i)}(t)\mathbb{1}_{t>\beta},$$

which is integrable on $\mathbb{R}_+$.
It suffices now to invoke Lemma 4.8 and Proposition 4.6 to deduce by dominated convergence that

$$\mathbb{P}_n^{(i)}(\pi_n(A)=k,\ T_{\mathrm{or}}\in B)\underset{n\to\infty}{\longrightarrow}\int_B \mathbb{P}(\pi^t(A)=k)\,h^{(i)}(t)\,\mathrm{d}t=\mathbb{P}(\pi^{(i)}(A)=k,\ \mathrm{e}_i\in B),$$

and this ends the proof. □

Proof of Corollary 4.2 :
Here we only prove a) since the proof of b) is identical. Fix $k\in\mathbb{N}$ and $A_1,\dots A_k$, Borel sets of $\mathbb{R}_+^*$ of zero Lebesgue measure boundary, satisfying $\sup A_i=\inf A_{i-1}$ for any $i\in\{2,\dots,k\}$. We set $B_i=(0,1)\times A_i$ for any $1\le i\le k$. Then

$$\begin{aligned}\mathbb{P}_n^{(\infty)}(T_{n,1}\in A_1,\dots,T_{n,k}\in A_k)=\ &\mathbb{P}_n^{(\infty)}(\pi_n(B_1)=1,\dots,\pi_n(B_k)\ge 1)\\ &\underset{n\to\infty}{\longrightarrow}\ \mathbb{P}(\pi(B_1)=1,\dots,\pi(B_k)\ge 1)=\ \mathbb{P}(T_1\in A_1,\dots,T_k\in A_k),\end{aligned}$$

where the convergence follows from Theorem 4.1. Furthermore, this result clearly still holds if the sets (A_i) satisfy $\sup A_i\le\inf A_{i-1}$ instead of $\sup A_i=\inf A_{i-1}$.
To obtain the result in the case where $A_1,\dots A_k$ are non necessarily pairwise disjoint sets, it suffices to to decompose $\cup_{i=1}^k A_i$ into a partition of disjoint Borel sets and to apply the same reasoning as above. Let us prove this in the simple case $k=2$:

$$\begin{aligned}&\mathbb{P}(T_{n,1}\in A_1,T_{n,2}\in A_2)\\ &\quad=\mathbb{P}(T_{n,1}\in A_1\cap A_2,T_{n,2}\in A_1\cap A_2)+\mathbb{P}(T_{n,1}\in A_1\cap A_2,T_{n,2}\in A_2\backslash A_1)\\ &\quad+\mathbb{P}(T_{n,1}\in A_1\backslash A_2,T_{n,2}\in A_1\cap A_2)+\mathbb{P}(T_{n,1}\in A_1\backslash A_2,T_{n,2}\in A_2\backslash A_1)\\ &\qquad=\mathbb{P}(\pi_n(A_1\backslash A_2)=0,\ \pi_n(A_1\cap A_2)\ge 2)\\ &\qquad+\mathbb{P}(\pi_n(A_1\backslash A_2)=0,\ \pi_n(A_1\cap A_2)=1,\ \pi_n(A_2\backslash A_1)\ge 1)\\ &\qquad+\mathbb{P}(\pi_n(A_1\backslash A_2)=1,\ \pi_n(A_1\cap A_2)\ge 1)\\ &\qquad+\mathbb{P}(\pi_n(A_1\backslash A_2)=1,\ \pi_n(A_1\cap A_2)=0,\ \pi_n(A_2\backslash A_1)\ge 1),\end{aligned}$$

and we conclude as above, using Theorem 4.1. □

Proof of Proposition 4.3 :
a) We base our reasoning on the fact that a Poisson point measure on $\mathbb{R}_+^*$ with intensity measure $\alpha x^{-2}\mathrm{d}x$ is the pushforward measure by the continuous mapping $x\mapsto x^{-1}$ of a Poisson process with parameter α^{-1}. Let ν be such a Poisson process. Then for any $a\in\mathbb{R}_+^*$, recalling that ρ_0 is an exponential variable with parameter α^{-1} and $\mathrm{e}_0=\rho_0^{-1}$ a.s.,

$$\mathbb{P}(T_1\ge a)=\mathbb{P}(\pi((0,1)\times(a,+\infty))\ge 1)=\mathbb{P}(\nu(0,a^{-1})\ge 1)=\mathbb{P}(\rho_0\le a^{-1})=\mathbb{P}(\mathrm{e}_0\ge a),$$

and hence $T_1\overset{\mathcal{L}}{=}\mathrm{e}_0$. A similar reasoning shows that for any $k\ge 1$, $A_1,\dots A_k$, Borel sets of $\mathbb{R}_+^*$ of zero Lebesgue measure boundary, satisfying $\sup A_i<\inf A_{i-1}$ for any $i\in\{2,\dots,k\}$,

$$\mathbb{P}(T_1\in A_1,\dots,\ T_k\in A_k)=\mathbb{P}(\mathrm{e}_0\in A_1,\dots,\mathrm{e}_{k-1}\in A_k).$$

As in the previous proof, the case where the sets (A_i) are non pairwise disjoint can be proved with the same reasoning, decomposing $\cup_{i=1}^k A_i$ into a partition of disjoint sets. We can then conclude that $(T_1,\dots,T_k)$ is distributed as $(\mathrm{e}_0,\dots,\mathrm{e}_{k-1})$.

b) In the same way, for any $t \in \mathbb{R}_+^*$, a Poisson point measure on $(0,t)$ with intensity measure $\alpha x^{-2}\mathrm{d}x\mathbb{1}_{(0,t)}(x)$ is the pushforward measure by the mapping $x \mapsto x^{-1}$ of the restriction to $(t^{-1}, +\infty)$ of a Poisson process with parameter α^{-1}. Then by definition of $\pi^{(i)}$, for any $a, b \in \mathbb{R}_+^*$,

$$\begin{aligned}\mathbb{P}(T_{\mathrm{or}}^{(i)} \geq b,\ T_1^{(i)} \geq a) &= \int_{a\vee b}^{+\infty} \mathbb{P}(T_1^t \geq a)\,\mathbb{P}(\mathrm{e}_i \in \mathrm{d}t)\\ &= \int_{a\vee b}^{+\infty} \mathbb{P}(\pi((0,1)\times(a,t)) \geq 1)\,\mathbb{P}(\mathrm{e}_i \in \mathrm{d}t)\\ &= \int_0^{a^{-1}\wedge b^{-1}} \mathbb{P}(\nu((u,a^{-1})) \geq 1)\,\mathbb{P}(\mathrm{e}_i^{-1} \in \mathrm{d}u)\\ &= \mathbb{P}(\nu((\mathrm{e}_i^{-1}, a^{-1})) \geq 1,\ \mathrm{e}_i^{-1} \leq b^{-1}).\end{aligned}$$

Now for any $i \geq 0$, e_i^{-1} is distributed as the $(i+1)$-th atom of ν, hence

$$\mathbb{P}(\nu((\mathrm{e}_i^{-1}, a^{-1})) \geq 1,\ \mathrm{e}_i^{-1} \leq b^{-1}) = \mathbb{P}(\mathrm{e}_i^{-1} \leq b^{-1},\ \mathrm{e}_{i+1}^{-1} \leq a^{-1}).$$

As a conclusion, we have $(T_{\mathrm{or}}^{(i)}, T_1^{(i)}) \overset{\mathcal{L}}{=} (\mathrm{e}_i, \mathrm{e}_{i+1})$. With a similar reasoning we obtain the equality in law, for any $k \geq 1$, between $(T_{\mathrm{or}}^{(i)}, T_1^{(i)}, \ldots, T_k^{(i)})$ and $(\mathrm{e}_i, \ldots, \mathrm{e}_{i+k})$. □

Proof of Theorem 4.4 :
We denote by $\bar{\pi}^{(i)}$ the random point measure obtained from π by removing its i largest atoms. By the restriction property of the Poisson point measures, conditional on $T_i = t$, we have $\bar{\pi}^{(i)} \overset{\mathcal{L}}{=} \pi^t$. Recalling from Proposition 4.3.(i) that $T_i \overset{\mathcal{L}}{=} \mathrm{e}_{i-1}$, for any $a \in \mathbb{R}_+^*$ and $k \geq 0$ we have

$$\begin{aligned}\mathbb{P}(\bar{\pi}^{(i)}((a,+\infty)) = k) &= \int_0^{+\infty} \mathbb{P}(\bar{\pi}^{(i)}((a,+\infty)) = k \,|\, T_i = t)\, h^{(i-1)}(t)\mathrm{d}t\\ &= \int_a^{+\infty} \mathbb{P}(\pi^t((a,t)) = k)\, h^{(i-1)}(t)\mathrm{d}t + \mathbb{1}_{k=0}\int_0^a h^{(i-1)}(t)\mathrm{d}t\\ &= \mathbb{P}(\pi^{(i-1)}((a,+\infty)) = k),\end{aligned}$$

where the last equality follows from the definition of the Cox process $\pi^{(i-1)}$. □

A Appendix

Proposition A.1. *For any $k \in \mathbb{N}$, $l \in \mathbb{Z}$ satisfying $k \geq l$, and $x \in \mathbb{R}_+$, we define*

$$I_{k,l}(x) := \int_0^t \frac{t^{k-l}}{(1+t)^k}\mathrm{d}t$$

Then we have

(a) *for $k \geq 0$,* $I_{k,0}(x) = \int_0^x \frac{t^k}{(1+t)^k}\mathrm{d}t = x - k\ln(1+x) + \sum_{j=1}^{k-1}\frac{(-1)^{j-1}}{j}\binom{k}{j+1}(1-(1+x)^{-j})$,

(b) *for $k \geq 1$,* $I_{k,1}(x) = \int_0^x \frac{t^{k-1}}{(1+t)^k}\mathrm{d}t = \ln(1+x) + \sum_{j=1}^{k-1}\frac{(-1)^j}{j}\binom{k-1}{j}(1-(1+x)^{-j})$,

(c) *for $k \geq 2$,* $I_{k,2}(x) = \int_0^x \frac{t^{k-2}}{(1+t)^k}\mathrm{d}t = \frac{1}{k-1}\left(\frac{x}{1+x}\right)^{k-1}$.

Proof :
Using the binomial theorem to expand $\frac{t^k}{(1+t)^k} = \left(1 - \frac{1}{1+t}\right)^k$, we get

$$I_{k,0}(x) = \sum_{j=0}^{k} \binom{k}{j} (-1)^j \int_0^x (1+t)^{-j} \mathrm{d}t, \quad \text{and} \quad I_{k,1}(x) = \sum_{j=0}^{k-1} \binom{k-1}{j} (-1)^j \int_0^x (1+t)^{-j-1} \mathrm{d}t,$$

which easily leads to (a) and (b). □

Proposition A.2. *For any $k \in \mathbb{N}$, $l \in \mathbb{Z}$ satisfying $k \geq l$, and any $x \in \mathbb{R}_+$, define*

$$J_{k,l}(x) := \int_x^\infty \frac{t^{k-l}}{(1+t)^k} \mathrm{d}t.$$

Then for any $t \in \mathbb{R}_+$, $J_{k,l}(t) < \infty$ if and only if $l \geq 2$. In this case we have

$$J_{k,l}(x) = \sum_{j=0}^{k-l} \frac{x^j}{(1+x)^{j+l-1}} \frac{(j+1)\dots(j+l-2)}{(k-1)\dots(k-l+1)}, \tag{9}$$

and in particular $J_{k,l}(0) = \frac{(l-2)!}{(k-1)\dots(k-l+1)} = \left[(l-1)\binom{k-1}{l-1}\right]^{-1}$.

Proof :

First for any $l \geq 2$ and $x > 0$, $J_{l,l}(x) = \int_x^\infty \frac{\mathrm{d}t}{(1+t)^l} = \frac{1}{l-1}(1+x)^{1-l}$.
For any $k \geq l \geq 2$ and $x \geq 0$, an integration by parts gives $J_{k,l}(x) = \frac{k}{k-l+1} J_{k+1,l} - \frac{x^{k-l+1}}{(k-l+1)(1+x)^k}$.
Then, assuming that $J_{k,l}(x) = \sum_{j=0}^{k-l} \frac{x^j}{(1+x)^{j+l-1}} \frac{(j+1)\dots(j+l-2)}{(k-1)\dots(k-l+1)}$, we obtain

$$\begin{aligned} J_{k+1,l}(x) &= \sum_{j=0}^{k-l} \frac{x^j}{(1+x)^{j+l-1}} \frac{(j+1)\dots(j+l-2)}{k\dots(k-l+2)} + \frac{x^{k-l+1}}{k(1+x)^k} \\ &= \sum_{j=0}^{k-l+1} \frac{x^j}{(1+x)^{j+l-1}} \frac{(j+1)\dots(j+l-2)}{k\dots(k-l+2)}, \end{aligned}$$

and (9) is then proved by induction on k. □

Chapter IV

Perspectives

The purpose of this chapter is to show some perspectives of Chapter III : In a joint work in preparation with G. Achaz and A. Lambert, we wish to shed light on the relevance of simple models as alternatives to the so-called *standard neutral model*, i.e. the Kingman coalescent model with Poissonian neutral mutations. Here, we aim at comparing the relative consistence with biological data of a branching population model based on Chapter III on the one hand, and of other comparable models based on the Kingman coalescent on the other hand.

The so-called *neutral theory of molecular evolution* [Kim84] assumes that the majority of mutations we observe are neutral (meaning that they do not affect the reproductive success), and as a consequence, implies that divergence is mainly due to an accumulation of neutral mutations. When analyzing sequence data, evolutionary biologists generally first compare their data to the predictions of the standard neutral model. Rejecting the neutral model allows them then to state that alternative hypotheses such as selection or demography have to be accounted for. This is done by using neutrality tests, which test in fact the goodness-of-fit of the standard model. A major part of these tests is based on the frequency spectrum (defined in Section 0.1.5), which motivates here our approach, that consists in comparing some alternative neutral models by testing their agreement with data through the expected site frequency spectrum.

We present here some preliminary results, where we intend to estimate the foundation time of some human subpopulations, using data from the *1000 human genomes* [C$^+$12]. We would like to stress here that all the numerical work (simulations, data handling, figures) has been achieved by Guillaume Achaz.

We propose four one-parameter models : first, we consider models based on the Kingman coalescent model. On the one hand, the first two ones belong to the class of the so-called *Kingman coalescent models in varying environment* [GT94, EMRS10] and are constructed as scaling limits of constant population size models with deterministic growth :

- **Model $\mathbf{K_{LIN}}$** : Kingman coalescent with linear growth,
- **Model $\mathbf{K_{BN}}$** : Kingman coalescent with a bottleneck.

On the other hand, we consider

- **Model $\mathbf{K_{COND}}$** : Kingman coalescent conditioned on its time to the most recent common ancestor.

The last model relies on the branching population model studied in Chapter III :

- **Model BD** : Critical birth-death model with fixed time of origin.

We estimate the parameter of each model using a least squares type method, based on the normalized expected site frequency spectrum. The latter can be expressed as a linear combination of the expected coalescence times [Wak09, (4.22)], so that the normalized expected site frequency spectrum is invariant through changes of time scale : only the branch length ratios of the sample genealogy will affect the estimated parameter. As a consequence, for the foundation times estimated from the different models to have a common sense, we have to choose a time scale shared by the four models. We choose as « universal » time scale the so-called *coalescent time scale*, where time is counted in units of N generations, with N the extant population size. This naturally provides a common time scale for our three first models, but requires some adaptation for the fourth one, where the population size is random, even at present time.

1 Models

We present here the four models we want to compare. In the first section, we explain models $\mathrm{K_{LIN}}$,$\mathrm{K_{BN}}$ and $\mathrm{K_{COND}}$. The second section is devoted to model BD. In particular, we discuss the number of parameters of the model, and the interpretation of the coalescent time scale in this model.
In all the models, time is counted backwards from now : present time is called time 0, and for any $t > 0$, an event that occurred t units of time in the past is said to have happened at time t.

1.1 Three one-parameter models based on the Kingman coalescent

Coalescent model with deterministic population growth

We give here a construction of the *coalescent model with deterministic population growth* (or *coalescent in a varying environment*) [GT94] as a scaling limit of the Wright-Fisher model, where we force a deterministic evolution of the population size. Just as the Kingman coalescent is the scaling limit of a whole class of constant population size models (including the Wright-Fisher model), note that the present model is also a scaling limit of a larger class of models.

Fix $N \in \mathbb{N}$. Let $(G_r) \in \mathbb{N}^{\mathbb{Z}_+}$ be the sequence of the population sizes at each generation : $G_0 = N$ is the population size at present time (labeled generation 0), and G_r is the population size at generation r, i.e. r generations into the past. The population dynamics is described as follows : for any $r \in \mathbb{Z}_+$, the G_r individuals of generation r choose their ancestor uniformly at random from the G_{r+1} individuals of generation $r + 1$. We then introduce a rescaled population size function

$$\gamma_N : \left\{ \begin{array}{ccc} \mathbb{R}_+ & \longrightarrow & \mathbb{R}_+ \\ u & \longmapsto & \frac{1}{N} G_{\lfloor Nu \rfloor} \end{array} \right. ,$$

where the population size is rescaled by a factor $1/N$ and time is counted in units of N generations, and we assume that there exists a function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\lim_{N \to \infty} \gamma_N(u) = \gamma(u), \text{ for any } u \in \mathbb{R}_+.$$

Finally, we consider the population model arising as the limit, as $N \to \infty$, of the population dynamics described above.

Hereafter we consider two particular cases of this model :

Model $\mathbf{K_{LIN}}$ (linear growth) : We assume that the population had a linear growth and was founded (in the coalescent time scale) t units of time ago, so that the model has t as a unique parameter. The population size function γ is thus a linear function such that for any $u \in \mathbb{R}_+$,

$$\gamma(u) = \left(1 - \frac{u}{t}\right) \mathbb{1}_{[0,t]}(u).$$

The construction as a scaling limit can be recovered by taking, for example, $G_r = \left\lfloor N - \frac{r}{t} \right\rfloor$ for any $r \leq Nt$, and $G_r = 1$ else. A graphical illustration is given in Figure 1.

Model $\mathbf{K_{BN}}$ (bottleneck) : The population is assumed to have reached instantaneously its current population size, t units of time ago. The model has again a unique parameter t, and the population size function is given for any $u \in \mathbb{R}_+$ by

$$\gamma(u) = \mathbb{1}_{[0,t]}(u).$$

This model arises as the scaling limit of the discrete model with, for example, $G_r = N$ for any $r \leq Nt$, and $G_r = 1$ else (see Figure 1).

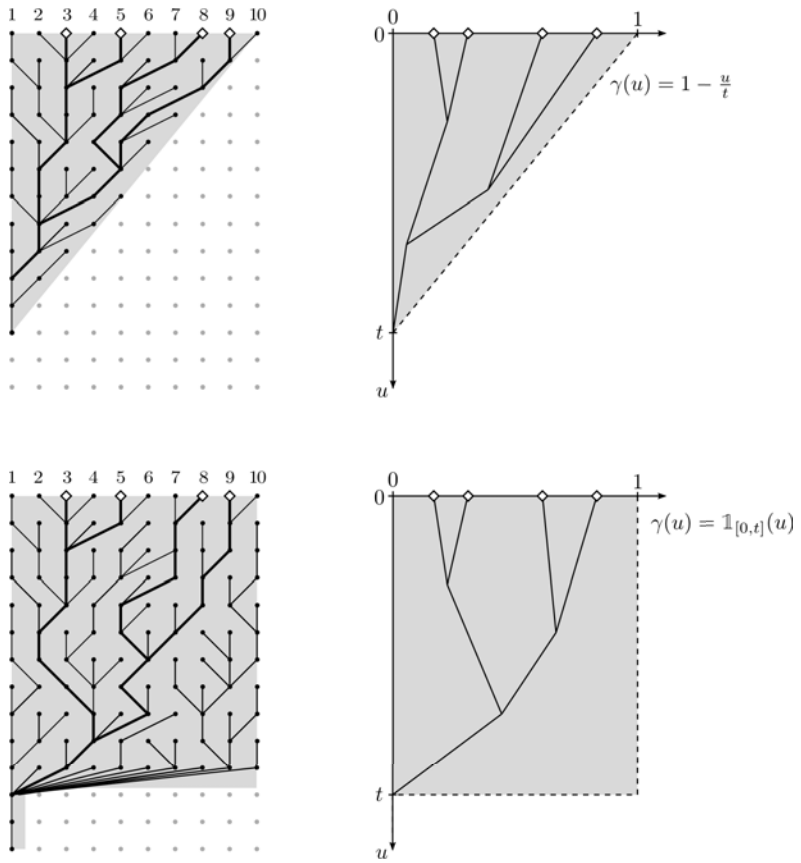

FIGURE 1 – On the left panel, a realization of a Wright-Fisher model with deterministic population growth (upper panel : linear growth ; lower panel : bottleneck), with $N = 10$ extant individuals at generation 0, and 4 sampled individuals (symbolized by $\diamond$). The genealogy of the sample appears in bold lines. On the right panel, a graphical representation of the genealogy for a sample of size 4 in the coalescent models $\mathrm{K_{LIN}}$ (upper panel) and $\mathrm{K_{BN}}$(lower panel), which can be seen as respective scaling limits of the models represented on the left panel.

Coalescent model conditioned on its time to the most recent common ancestor

We propose then a third model based on the Kingman coalescent. Let $\mathbb{P}_K$ denote the law of the Kingman coalescent for a (uniform) sample of size n, and let T_{mrca} denote the time to the most recent common ancestor of the sample. Then the genealogy of a sample of size n in a population originating at time t, in model K_{COND}, is defined by $\mathbb{P}_K(\cdot \mid T_{\text{mrca}} \leq t)$.

Of course, the conditioning on $T_{\text{mrca}} \leq t$ does not ensure that the time to the most recent common ancestor of the whole population is itself lower than t. In practice however, for samples of sufficient size (such as ours, whose size is about several hundreds), the time to the most recent common ancestor of the sample is equal to that of the whole population with high probability. Hence, model K_{COND} with parameter t can be seen as a Kingman coalescent model in which the time to the most recent common ancestor of the whole population is lower than t. We admit that its interpretation as a population model with foundation time t is not obvious, but we intend in the future to investigate the way populations are growing in the different models, in particular to answer this question.

1.2 Critical branching population model

The fourth model we propose is based on the model described in Chapter III and is called **BD model** (where BD stands for birth-death). In the three models based on the Kingman coalescent, the assumption of deterministic population size provides a common time scale (the coalescent time scale) in which the parameter t naturally makes sense as the foundation time of the population. On the contrary, because of its randomly fluctuating population size, the choice of a counterpart to the coalescent time scale for model BD is less obvious. Besides, this model has more than one parameter, while we wish to construct a one-parameter model. Hence, after having first recalled the population dynamics of the model, we second discuss the interpretation of the parameters, and we finally explain how the final (one-parameter) model BD is defined.

Population dynamics

We start with the model constructed in Section 1.1. This model has (besides the sample size n) three parameters, but only two degrees of freedom with respect to the law of the sample genealogy. For the sake of simplicity we consider here a two-parameter version of the model of Chapter III. We recall, according to Chapter III, that this model is also the scaling limit of a larger class of branching population models.

Fix $n \in \mathbb{N}$, $T \in \mathbb{R}_+$ and $p \in (0,1)$. Consider a critical birth-death tree with birth and death rates both equal to 1, started at time T. Assume that any individual alive at present time is independently sampled with probability p, and condition the tree on having n sampled individuals. The n sampled individuals are then *a posteriori* uniformly distributed among the extant population.

Discussion on the parameters

From Theorem III.2.1, we deduce that the coalescent point process of the sample is distributed as a sequence of $n-1$ i.i.d. random variables with probability density function $x \mapsto (1+x)^{-2}$, conditioned to be smaller than $\tau := pT$, and then rescaled by a factor p^{-1}. Thus, the effects of the two parameters p and T on the law of the genealogy are twofold, since they affect both the branch length ratios and the time scale. However, for our purpose, only the effect on the branch length ratios matters. Indeed, our estimations are later based on the normalized site frequency

spectrum, which can be expressed as a linear combination of the expected branching times, and hence the estimation is independent of the choice of the time scale.

From this point of view, the model reduces then to one parameter, namely τ. Indeed, fix τ and consider $p, p' \in (0,1)$ and $T, T' \in \mathbb{R}_+$ satisfying $pT = p'T' = \tau$. Then the sample genealogy in the model with parameters (p,T) has the law of the sample genealogy in the model with parameters (p',T') up to a rescaling of time by a factor p/p'. In particular, for any pair (p,T) such that $pT = \tau$, the model produces the same renormalized expected site frequency spectrum. Note that this can also be seen directly from the formula provided by Proposition III.3.1.

A coalescent time scale in the BD model.

In the three models $\mathrm{K_{LIN}}$, $\mathrm{K_{BN}}$ and $\mathrm{K_{COND}}$, time is counted in units of N generations, where N is the total extant population size. This time scale has no obvious meaning in the BD model, since its population growth, and hence its extant population size, are random. We explain here how to give a meaning to the coalescent time scale in the BD model.

The question can be reformulated as follows : when estimating the parameter τ in the model described above, which quantity provides a measure of the foundation time of the population in the coalescent time scale ? In the BD model, we start with a birth-death process with parameter 1 and foundation time T, so that one unit of time corresponds on average to one generation, and the absolute foundation time of the population is thus T. Hence a measure of the foundation time in the coalescent time scale would be given by a rescaling of time T by the total extant population size.

Let us now consider a realization of the BD model and denote by N its (realized) extant population size (recall that the total extant population size is not an observed data). The time of foundation in the coalescent time scale, which we wish to estimate, is thus given by T/N. We denote by $\hat{\tau}$ an estimator of the parameter $\tau = pT$ (using a least squares type method specified later), so that an estimator of T/N is then $\frac{\hat{\tau}}{pN}$. Now since n individuals are sampled from the total extant population, each of them having a probability p to be sampled, we can estimate the parameter p by its maximum likelihood estimator $\hat{p} = n/N$. As a conclusion, an estimator of the foundation time of the population is $\hat{\tau}/n$. From now on, we replace thus the parameter τ of model BD by the parameter $t = \tau/n$, so that in all four models $\mathrm{K_{LIN}}$, $\mathrm{K_{BN}}$, $\mathrm{K_{COND}}$ and BD, estimating their parameter t means estimating the foundation time of the population in the common coalescent time scale.

2 Estimation of a foundation time and comparison between the models

2.1 Simulations and numerical computation

Let us first present the sequence data we use, namely public datasets from the 1000 human genomes [C$^+$12]. In total, 1092 human genomes have been sequenced. The files we have access to describe in particular the location and allele frequency of the observed single nucleotide polymorphisms (SNP). Table 1 is a simplified version of one of the files provided by [C$^+$12] : it shows, for 9 of the hundreds of listed SNP's in chromosome 20, the alleles carried by 10 of the 1092 human genomes. Besides, the sequenced human genomes are classified into subpopulations such as African populations, Asian population, etc. Obtaining the site frequency spectrum of a sample

of size n among theses genomes is quite simple : it suffices to count the number of mutated alleles carried by k individuals, $1 \leq k \leq n-1$. In the example given by Table 1, all the mutations are in the heterozygous state, except for HG096, which is homozygous for the mutation at position 61795. Hence, if we assume that the reference allele is the ancestral allele (the choice of the ancestral allele is later discussed), two mutations (SNP's 60828 and 62255) are carried by one individual ($\xi_1 = 2$), 1 mutation (SNP 61098) is carried by 3 individuals ($\xi_3 = 1$) and 1 mutation (SNP 61795) is carried by 5 individuals, one of them being homozygous so that $\xi_6 = 1$. Note that

#CHROM	POS	REF	ALT	HG096	HG097	HG099	HG100	HG101	HG102	HG103	HG104	HG106	HG108	...
20	60479	C	T	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	...
20	60522	T	TC	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	...
20	60571	C	A	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	...
20	60828	T	G	0|0	0|1	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	...
20	61098	C	T	0|0	0|0	0|0	0|0	0|1	0|0	0|1	0|1	0|0	0|0	...
20	61279	C	T	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	...
20	61795	G	T	1|1	0|0	0|0	0|0	0|1	0|0	0|1	0|1	0|1	0|0	...
20	62100	T	C	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	0|0	...
20	62255	T	C	0|0	0|0	0|0	0|1	0|0	0|0	0|0	0|0	0|0	0|0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

TABLE 1 – Example of (simplified) data from the 1000 human genomes.
`#CHROM` : Chromosome number.
`POS` : Position of the SNP in the chromosome.
`REF` : Reference allele (the first sequenced human genome is arbitrarily considered as reference).
`ALT` : Alternate allele(s). At position 60522, there is in fact an insertion instead of a SNP : here the reference allele T is followed by an additional DNA base C in the genomes carrying this mutation.
`HG096` : The two values $x_1|x_2$ of column HG096 indicate the allele carried by each of the two independent copies of chromosome 20 in human genome number 096. Value $x = 0$ means that the carried allele is the reference allele, $x = 1$ means that it is one of the alternate alleles `ALT`.

Let us now describe the way we obtained the expected frequency spectra in the different models : apart from the BD model, for which Proposition III.3.1 provides an explicit formula for the expected site frequency spectrum, the latter is computed numerically for models $\mathrm{K}_{\mathrm{LIN}}$, K_{BN} and $\mathrm{K}_{\mathrm{COND}}$, using simulations. As far as models $\mathrm{K}_{\mathrm{LIN}}$ and K_{BN} are concerned, analytic results for the expected coalescence times in the Kingman coalescent models with deterministic growth are provided by [EMRS10], but even in the simple cases of models $\mathrm{K}_{\mathrm{LIN}}$ and K_{BN}, the obtained formulae are not easily numerically computable.

Models $\mathbf{K_{LIN}}$ and $\mathbf{K_{BN}}$: In coalescent models with varying environment, the evolution imposed to the population size can be interpreted as changes in the coalescence rate over time. Indeed, define for all $u > 0$, $G(u) = \int_0^u \frac{\mathrm{d}v}{\gamma(v)}$, where γ denotes the population size function. let $\mathcal{T}$ be a Kingman coalescent for a sample of size n ; conditional on its node depths $t_2 > \ldots > t_n$, define by $\mathcal{T}_\gamma$ the coalescent with n leaves and node depths $(G^{-1}(t_i))_{2 \leq i \leq n}$. Then $\mathcal{T}_\gamma$ has the law of a coalescent with population size function γ. In our particular cases $\mathrm{K}_{\mathrm{LIN}}$ and K_{BN}, the function G^{-1} has a simple expression which makes simulations quite fast : $\mathcal{T}_\gamma$ can be simply obtained by applying G^{-1} to the node depths of the Kingman coalescent tree $\mathcal{T}$.

Model $\mathbf{K_{COND}}$: Simulations are done by a rejection algorithm. This is of course time-consuming, especially for small values of t.

2.2 Estimation method

When a site on a DNA sequence is polymorphic, one cannot always determine on a reliable basis which of the alleles is the ancestral one. For example, if a site is dimorphic, making the wrong choice for the ancestral allele leads to decide that the mutated one is carried by k individuals, while it is in fact carried by $n-k$ individuals. To avoid such errors, let us define the folded site frequency spectrum $(\eta_k)_{1\leq k\leq \lfloor \frac{n}{2} \rfloor}$ by

$$\eta_k = \frac{\xi_k + \xi_{n-k}}{1+\delta_{i,n-i}},$$

where $(\xi_k)_{1\leq k\leq n-1}$ is the site frequency spectrum of the sample (of size n) and $\delta_{i,j}=1$ iff $i=j$, else $\delta_{i,j}=0$. For any $1\leq k\leq \lfloor \frac{n}{2} \rfloor$, η_k is thus the number of polymorphic sites carried by both k and $n-k$ individuals.

The estimation method we use is a least-squares type method, based on the following distance, defined for any $t>0$ by

$$d(t) = 2\sum_{k=3}^{n-1} \frac{(\mathbb{E}^t(\eta_k) - \eta_k^{\text{obs}})^2}{\mathbb{E}^t(\eta_k) + \eta_k^{\text{obs}}},$$

where (η_k^{obs}) denotes the folded site frequency spectrum of the observed data, and with a slight abuse of notation, $\mathbb{E}^t$ denotes the law of the sample genealogy under any of our models with parameter t. In the definition of the distance we deliberately ignored η_1 and η_2. Indeed, sequence data are damaged by frequent sequencing errors (i.e. errors made during « reading » the DNA sequences, which result in point substitutions of one DNA base by another), that mostly affect singletons and doublons (mutations carried by one or two individuals).

The estimated foundation time is then defined by the value of t that minimizes distance d. In order to determine this minimal value, we scan the time interval $[0.8, 10]$ by increments of 0.01. Values smaller than 0.8 are not taken into account, in particular because they require too long numerical computations in case K_{COND}. Besides, for the three models K_{LIN}, K_{BN} and K_{COND}, for any considered value of $t\in[0.8,10]$, the computation of distance $d(t)$ relies on 10^5 simulation replicates.

2.3 Results and conclusion

Figure 2 illustrates the results for the estimation of the foundation time under our four models for two examples of human populations : African populations on the one hand, and European populations on the other hand.
The upper graphs show distance d as a function of the parameter t, for all four models K_{LIN}, K_{BN}, K_{COND} and BD.
Then, given the estimated foundation time T_f, the lower graphs represent the expected « weighted folded site frequency spectrum » $(\theta_k)_{1\leq k\leq \lfloor \frac{n}{2} \rfloor}$ defined below, and normalized to 1 on the graphs ([Ach09]). For any $1\leq k\leq \lfloor \frac{n}{2} \rfloor$,

$$\theta_k = \frac{k(n-k)}{n}\,(1+\delta_{k,n-k})\,\eta_k.$$

We choose to draw $(\mathbb{E}(\theta_k))$ rather than $(\mathbb{E}(\eta_k))$ because it provides an easy comparison with the Kingman coalescent, for which $(\mathbb{E}(\theta_k))$ is constant : we know that the expected site frequency spectrum of the Kingman coalescent is given by $\mathbb{E}(\xi_k)=\theta/k$ [Wak09, (4.20)], which leads to

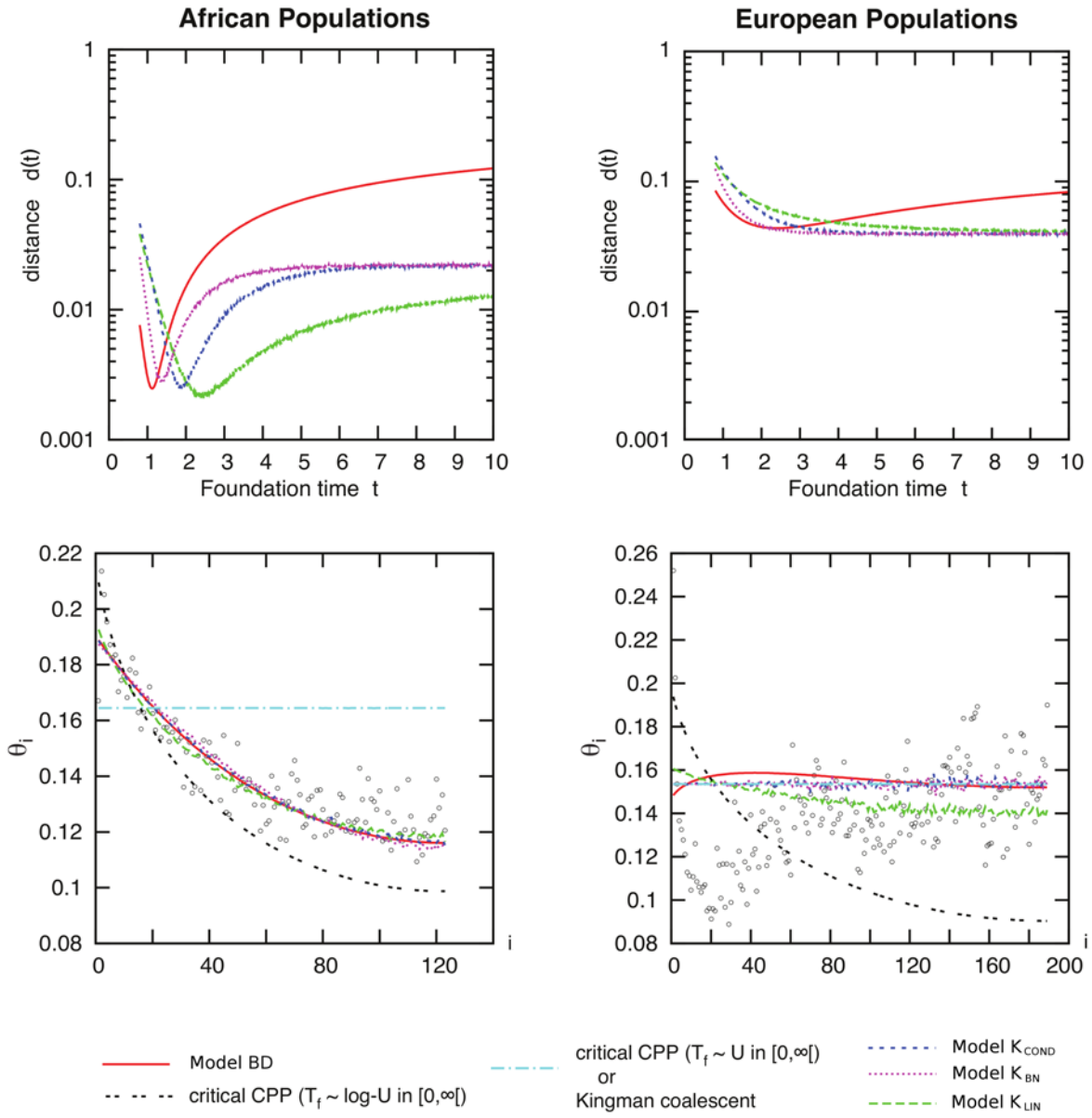

FIGURE 2 – Estimation of the foundation time under models K_{LIN}, K_{BN}, K_{COND} and BD, for African populations (left panel) and European populations (right panel). (Colored figure in the electronic version)
Upper panels : Distance d as a function of the parameter t, between observed and predicted site frequency spectra. The estimated foundation time T_f of a given model is given by the value of t minimizing the corresponding distance.
Lower panels : Estimated frequency spectrum (θ_i) (corresponding to the estimated foundation time T_f) in dotted lines. Observed data are represented by circles.

$\mathbb{E}(\eta_k) = \left(\frac{\theta}{k} + \frac{\theta}{n-k}\right)/(1+\delta_{k,n-k})$, and hence $\mathbb{E}(\theta_k) = \theta$. The normalization by $\sum \mathbb{E}(\theta_k)$ also serves the purpose of comparison between the different models.

For each estimated frequency spectrum obtained by simulation (models K_{LIN}, K_{BN}, K_{COND}), the plots are based on 10^6 replicates. We also added the plots of the expected θ_k given by two zero-

parameter models, namely the Kingman coalescent (or equivalently, the birth-death model with uniform prior on its time of origin, see Proposition III.3.4), for which $\mathbb{E}(\theta_k)$ is constant, and the birth-death model with log-uniform prior on the foundation time (see Proposition III.3.5).

In the case of African subpopulations, the lower graph shows that the estimated expected site frequency spectra for models K_{LIN}, K_{BN}, K_{COND} and BD are in excellent agreement with the observed data. Besides, these models give obviously a really better fit than the two zero-parameter models given by the Kingman coalescent and the critical birth-death model with log-uniform prior.

However, although the estimated expected site frequency spectra in all four models K_{LIN}, K_{BN}, K_{COND}, and BD are very similar, the estimated foundation times T_f, specified below, are significantly different.

Model	BD	K_{BN}	K_{COND}	K_{LIN}
T_f	1.12	1.34	1.86	2.39

In the case of European subpopulations, it is clear that none of our one-parameter models properly fits the observed frequency spectrum, since they all fail in particular to explain the fall and rise of the site frequency spectrum at intermediate values of $k \approx 10 - 30$. It seems reasonable to think that no one-parameter model based on the Kingman coalescent or on the birth-death model would be able to fit these data. We chose here not to present results concerning American or Asiatic populations, since the conclusions are very similar to those obtained for European populations.

In the light of these results, a first conclusion would be that there is no reason to give priority to one of the models in particular. On the other hand, we admit that the common time scale providing our interpretation of the parameter t in the four models as foundation time of the population remains questionable, and that the differences between the respective estimated foundation times may arise from this. Note however that there is no doubt about the definition of t as a foundation time in models K_{LIN} and K_{BN}, and that these two models give though two really distinct estimated foundation times (respectively 2.39 and 1.34).

Second, we would like to say a word about the comparison between the results obtained for European and African subpopulations. As a matter of fact, a one-parameter model seems sufficient to fit the observed data in the African case, but not in the European case. This reflects a population growth easier to model in the African subpopulation than in the case of European subpopulations. Our results are hence consistent with the known fact that in the modern human history, African populations, which appeared earlier and are at the origin of the other populations, have a quite simple evolution with a regular growth, compared e.g. to European populations, whose history is far more recent and intricate.

Finally, a natural question is the meaning of the estimated foundation times in the real time scale : can we evaluate these coalescent times in years ? Our first, basic answer to this question is not convincing. Let us make a simple calculation : to convert the estimated foundation times into years, we first need to estimate the parameter N. This parameter describes the extant population size : let us take 10^8 as a reference value for the « extant »population size of African populations, which corresponds to an order of magnitude of this size between years 1750 and 1950 [Uni10] - the recent explosion of demography is on purpose not taken into consideration, since it probably has no effect on divergence of sequences yet. This being said, we should in fact

only take into account the individuals in the extant population that currently play a role in the demography : this is known in population genetics as the *effective* population size. A ratio of order of magnitude 1/10 seems reasonable to estimate the proportion of the population really involved in the demography. It yields thus $N = 10^7$. Besides, the order of magnitude usually used for a generation is 20 years. Hence, our estimated foundation times (which are all around 1 in the coalescent time scale) would give an estimated foundation time for the African subpopulations of around $20N = 2.10^8$ years. This is of course not reasonable in the light of the current knowledge on the evolutionary history (200 million years is known to correspond to the age of dinosaurs).

The question of the interpretation of the coalescent time scale is thus still open. Besides, since the goodness-of-fit of a model with data does not necessarily ensure the relevance of the model, we also mean to investigate the population growth in the models we are working with, in order to understand, if so, why they could not reasonably model the demographic history of human subpopulations.

Bibliography

[Ach09] G. Achaz. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–258, 2009.

[ADL14] G. Achaz, C. Delaporte, and A. Lambert. Sample genealogy and mutational patterns for critical branching populations. *in prep.*, 2014.

[Ald93] D. Aldous. The continuum random tree III. *The Annals of Probability*, pages 248–289, 1993.

[AP05] D. Aldous and L. Popovic. A critical branching process model for biodiversity. *Advances in applied probability*, 37(4):1094–1115, 2005.

[Ber91] J. Bertoin. Sur la décomposition de la trajectoire d'un processus de Lévy spectralement positif en son infimum. In *Annales de l'IHP Probabilités et statistiques*, volume 27, pages 537–547. Elsevier, 1991.

[Ber92] J. Bertoin. An extension of Pitman's theorem for spectrally positive Lévy processes. *The Annals of Probability*, 20(3):1464–1483, 1992.

[Ber96] J. Bertoin. *Lévy processes*, volume 121. Cambridge university press, 1996.

[Ber10] J. Bertoin. A limit theorem for trees of alleles in branching processes with rare neutral mutations. *Stochastic Processes and their Applications*, 120(5):678–697, 2010.

[C+12] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[CD10] L. Chaumont and R.A. Doney. Invariance principles for local times at the maximum of random walks and Lévy processes. *The Annals of Probability*, 38(4):1368–1389, 2010.

[CL12a] N. Champagnat and A. Lambert. Splitting trees with neutral Poissonian mutations I: Small families. *Stochastic Processes and their Applications*, 122(3):1003–1033, 2012.

[CL12b] N. Champagnat and A. Lambert. Splitting trees with neutral Poissonian mutations II: Largest and Oldest families. *Stochastic Processes and their Applications*, 2012.

[CLR12] N. Champagnat, A. Lambert, and M. Richard. Birth and death processes with neutral mutations. *International Journal of Stochastic Analysis*, 2012, 2012.

[CM68] K.S. Crump and C.J. Mode. A general age-dependent branching process. I. *Journal of Mathematical Analysis and Applications*, 24(3):494–508, 1968.

[Del13a] C. Delaporte. Lévy processes with marked jumps I: Limit theorems. *Journal of Theoretical Probability*, 2013.

[Del13b] C. Delaporte. Lévy processes with marks II : Application to a population model with mutations at birth. *Eprint arXiv:1305.6491*, 2013.

[DL02] T. Duquesne and J-F. Le Gall. *Random trees, Lévy processes and spatial branching processes*, volume 281. Société mathématique de France, 2002.

[DN70] H.A. David and H.N. Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[Don07] R.A. Doney. *Fluctuation theory for Lévy processes: École D'Été de Probabilités de Saint-Flour XXXV-2005*, volume 1897. Springer, 2007.

[Dur08] R. Durrett. *Probability models for DNA sequence evolution*. Springer, 2008.

[EMRS10] A. Eriksson, B. Mehlig, M. Rafajlovic, and S. Sagitov. The total branch length of sample genealogies in populations of variable size. *Genetics*, 186(2):601–611, 2010.

[Ewe72] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112, 1972.

[Gei96] J. Geiger. Size-biased and conditioned random splitting trees. *Stochastic processes and their applications*, 65(2):187–207, 1996.

[Ger08] T. Gernhard. New analytic results for speciation times in neutral models. *Bulletin of mathematical biology*, 70(4):1082–1097, 2008.

[GK97] J. Geiger and G. Kersting. Depth-First Search of Random Trees, and Poisson Point Processes. In *Classical and modern branching processes*, pages 111–126. Springer, 1997.

[GT94] R.C. Griffiths and S. Tavare. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1310):403–410, 1994.

[Jag69] P. Jagers. A general stochastic model for population development. *Scandinavian Actuarial Journal*, 1969(1-2):84–103, 1969.

[JS87] J. Jacod and A.N. Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer-Verlag Berlin, 1987.

[Kal02] O. Kallenberg. *Foundations of modern probability*. Springer, 2002.

[Kar75] A.F. Karr. Weak convergence of a sequence of Markov chains. *Probability Theory and Related Fields*, 33(1):41–48, 1975.

[Kim69] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[Kim84] M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1984.

[Kin82a] J.F.C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[Kin82b] J.F.C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[Kyp06] A. Kyprianou. *Introductory lectures on fluctuations of Lévy processes with applications*. Springer, 2006.

[Lam67] J. Lamperti. The limit of a sequence of branching processes. *Probability Theory and Related Fields*, 7(4):271–288, 1967.

[Lam08] A. Lambert. The allelic partition for coalescent point processes. *Markov Proc. Relat. Fields. 15 359-386.*, 2008.

[Lam10] A. Lambert. The contour of splitting trees is a Lévy process. *The Annals of Probability*, 38(1):348–395, 2010.

[Lam11] A. Lambert. Species abundance distributions in neutral models with immigration or mutation and general lifetimes. *Journal of mathematical biology*, 63(1):57–72, 2011.

[LS12] A. Lambert and F. Simatos. Asymptotic behavior of local times of compound Poisson processes with drift in the infinite variance case. *Journal of Theoretical Probability*, pages 1–51, 2012.

[Nag64] M. Nagasawa. Time reversions of Markov processes. *Nagoya Mathematical Journal*, 24:177–204, 1964.

[OP09] J. Obłój and M. Pistorius. On an explicit Skorokhod embedding for spectrally negative Lévy processes. *Journal of Theoretical Probability*, 22(2):418–440, 2009.

[Pop04] L. Popovic. Asymptotic genealogy of a critical branching process. *Annals of Applied Probability*, pages 2120–2148, 2004.

[Ric14] M. Richard. Splitting trees with neutral mutations at birth. *Stochastic Processes and their Applications*, 124(10):3206–3230, 2014.

[Sta08] T. Stadler. Lineages-through-time plots of neutral models for speciation. *Mathematical biosciences*, 216(2):163–171, 2008.

[Sta09] T. Stadler. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66, 2009.

[Taï92] Z. Taïb. *Branching processes and neutral evolution*. Springer-Verlag, 1992.

[Uni10] United Nations Population Division. The World at Six Billion. http://www.un.org/esa/population/publications/sixbillion/sixbilpart1.pdf, 2010.

[Wak09] J. Wakeley. *Coalescent theory: an introduction*. Roberts & Company, 2009.